Using machine learning systems to investigate phonological representations

A Dissertation Presented

by

Kalina Kostyszyn

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Doctor of Philosophy

in

Linguistics

Stony Brook University

August 2024

Stony Brook University

The Graduate School

Kalina Kostyszyn

We, the dissertation committee for the above candidate for the
Doctor of Philosophy degree, hereby recommend
acceptance of this dissertation.

Jeffrey Heinz — Dissertation Advisor
Professor, Department of Linguistics and Institute for Advanced Computational Science

Marie K. Huffman — Chairperson of Defense
Professor, Department of Linguistics

Owen Rambow
Professor, Department of Linguistics and Institute for Advanced Computational Science

Susan E. Brennan
Distinguished Professor, Department of Psychology

This dissertation is accepted by the Graduate School

Celia Marshak
Dean of the Graduate School

Abstract of the Dissertation

Using machine learning systems to investigate phonological representations

by

Kalina Kostyszyn

Doctor of Philosophy

in

Linguistics

Stony Brook University

2024

Language learning is a complex issue of interest to linguists, computer scientists, and psychologists alike. While the different fields approach these questions at different levels of granularity, findings in one field profoundly affect how the others proceed. My dissertation examines the perceptual and linguistic generalizations regarding the units that make up words (phonemes, morphemes, and vocal quality) in Polish and English to better understand how both humans and computers formulate these concepts in language.

I use computational modeling and machine learning to investigate Polish morphophonology in two ways. First, I examine consonant clusters at the beginning of Polish words to see what parameters determine human-like learnability, compared to a survey of native speakers. I run several studies to compare learning with gradient or categorical data, each at the cluster, bigram, and featural level. Second, I examine Polish yer alternation and study whether machine learning approaches can generalize morphophonological information to target this pattern when given a larger Polish. Using low level neural networks and a classification-and-regression tree (CART) decision algorithm, I examine how well they use morphological and phonological information to make generalizations that capture a small subset of the Polish vocabulary.

Additionally, I conduct a psycholinguistic experiment with English speakers to further establish what level of attention listeners may give when building phonological representations. I test this by extending a previous study finding that real word primes make rejection of nonword primes more difficult, determining that the effect generalizes across speakers.

This research addresses a tension in modeling the computational problem of language learning between the formalization of representation and the mechanics of the learning apparatus. Different levels of abstraction can give more sophisticated insight into the data at hand, but at a cost that may not be representative of human learning. I argue that computational linguistic questions such as these provide an interesting window into the strengths and limitations of machine learning questions as compared to the human language learning faculty.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

People say many contradictory wonderful and terrible things about working on a doctorate, and all of them are true. But what makes the experience worthwhile is the amazing conga line of supporters and friends that you meet along the way.

The obvious place to start is thanking my advisor Jeffrey Heinz, without whom none of this would be possible. Jeff, I will always envy your sheer passion for knowledge, and I hope you understand how much we as students value the energy you bring to the table! You've been invaluable in shaping this dissertation, and my trajectory through graduate school as a whole. I can only hope to lead others as you have!

My committee was made up of tremendous guiding forces. Marie, your voice and approach to teaching is a huge inspiration to me, and you'll never understand how much I appreciated your help when I instructed a course of my own. Owen, I greatly value the experience you bring the the department and how you interface with computer scientists and psychologists to bring everyone into the conversations. Susan, working with you in the BIAS-NRT program has been an amazing way to expand my boundaries, and I could never begin to imagine being part of the projects the group has taken on.

The faculty of Stony Brook's Linguistics department as a whole went above and beyond in looking out for its students, especially in its response to the quarantines for COVID-19. When people from other universities and departments come into the fold, they are always astounded by how caring and congenial you all are. I can only hope to make you all proud. I have to thank Lori Repetti in particular, for looking out for us as new graduate students, both in her capacity as department chair and as a generally deeply caring and welcoming person. I also thank Thomas Graf, who has been instrumental in shaping the department community and who reached across the phonological-syntactic divide to take me on as a research assistant.

I've met and worked with phenomenal linguists during my time at Stony Brook, and I wish you all the best as our paths diverge! To Alëna Aksenova, Aniello De Santo, Pardis Derakhshandeh, Hossep Dolatian, Félix Fonseca, Daniel Greeson, Salam Khalifa, Dakotah

Lambert, Han Li, Magdalena Markowska, Scott Nelson, Sarah Payne, Andrija Petrović, Jon Rawski, John David Storment, Logan Swanson, Neda Taherkhani, Anastasiia Voznesenskaia, Chia-Chi Yu, and so many more - working and spending time with you all brightened each day and made the work a little easier!

Outside of linguistics, peers around Stony Brook helped me look outside of my box and see how all our individual pieces fit together. To Arthur Samuel, thank you for taking me on in your lab and giving me opportunity to learn a broad new range of skills. To my BIAS-NRT friends - Carl Wiedemann, Rosa Bermejo, Mackenzie Johnson, John Murzaku, Amie Paige, Adil Soubki - our conversations are always enlightening, and I wish you all success in what comes next!

There's a special place in my heart for all my friends outside of Stony Brook who supported me with book clubs, D&D, and traipsing around New York. Juhi Aggarwal, Rebecca Kaplan, Christina Pellegrino, Rachel Sadaty-Ellerson, Maddie Smith, and Nora Broderick - you're all exceptionally talented people, and I'm glad to have kept you around so long. I also thank Tabatha Barton-Nypower, Kimberly Landstrom, and Connie Wen, who have been some of my closest friends for years and put up with more than I can express here.

And finally to my family. To my maternal grandfather, Wojciech Michalski, who passed the summer before I could complete my doctorate, my maternal grandmother Kazimiera Michalska, and my paternal grandmother, Lucyna Kostyszyn. My younger brother, Olek Kostyszyn. And my parents, Marek and Agnieszka. Nigdy nie zrozumiecie, jak bardzo was wszystkich kocham.

This material is based upon work supported by the National Science Foundation under Grant No. NRT-HDR 2125295. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

1: INTRODUCTION

This dissertation examines the tension of sparse versus rich representations through three case studies, two computational problems from different areas of phonology and one psycholinguistic problem related to word perception and access. For the computational problems, I examine the effectiveness of different modeling techniques and what can be learned from the successes and failures of each. I investigate level of detail, frequency, and abilities of these models to make generalizations that represent one phonotactic and one morphophonological question. For the psycholinguistic question, I study the effect of multiple speakers with different voices on recent memory to investigate differences in representation that can attributed to different levels of sound processing. These studies invoke a larger discussion among phonologists about how having a variety of methods to solve narrow problems can better inform how these processes interact at higher levels, and what lessons can be exchanged with computer scientists and psychologists about these conclusions.

Richer representations can give more sophisticated insight into the data at hand, but may be less resilient to variation between speakers and contexts. On the other hand, certain model architectures may be better suited for more basic and austere representations. Probablistic models may have difficulty distinguishing low-frequency patterns without sufficient information to form a generalization. Here, I look at the tension between these positions, and why certain representations may be more appropriate for different tasks.

A key aspect for developing phonological representations is determining what information is needed for a given task. Sapir (1925) highlights two points of variation, at the *speaker* level and at the *language* level, which push and pull on one another to create the speech sounds and patterns of a language. On the speaker level, phonetic differences expand the boundaries of phonemic categories to accommodate variation in pitch, volume, and so on. Meanwhile, at the language-level, those boundaries nonetheless restrict what is in a language and what is not. Later studies strengthened this by showing it extended not only to language acceptability, but even language *perception* (Dupoux et al., 1998), with failures to perceive sound sequences that were not possible to produce in a speaker's known language. This disparity has called into question how this information is stored, such that certain speech sounds are accepted and understood while others are rejected.

More recently, Pierrehumbert (2016) highlights the relationship between the abstract accumulation of information about speech and the phonetic realization of that information. Richer representations of phonology incorporate this breadth of information at multiple levels and express them according to the problem at hand. However, there are arguments against these highly detailed representations, suggesting instead that the focus of linguistics should be on finding singular, minimalist answers in search of "one correct grammar" (Hyman, 1970). Pierrehumbert argues against this, noting that highly contextual information can influence variation in speech and thus should be considered in theory.

This is not to say that minimalist theories with sparse detail do not contribute to theory. Rather, they are only one approach, and while they may sufficiently answer some questions, more detail may be required to answer others. Sapir (1925) warns against using phonetic methods to explain phonological problems, but there are time when integrating the two provides a richer explanation to a question. For example, Huang et al. (2018) investigate effects of prosodic focus on Mandarin tone and find not only that creaky phonation can be used to distinguish tones, but also that creakiness is used to signal focus despite not being phonemic in Mandarin. Moreover, they add that the male participants generally used more creaky phonation than female participants. They attributed this to the male speakers' generally deeper voices limiting their F0 range, necessitating a secondary phonetic cue. The use of prosodic experiments uncovers a more nuanced view of the phonetic space and phonological goals of these speaker. Sapir warns to not (emphasis mine) "*confuse linguistic structure with a particular method* of studying linguistic phenomena," highlighting that any single approach can give limited insight to the true scope of a phenomenon. Minimalist theories may seek simpler representations, but in doing so, lose the detail that a richer grammar allows - certainly, it is redundant that a simpler representation will have simpler details. Embracing a richer grammar, on the other hand, allows one to look at linguistic problems from different angles and tease out details lost from a more abstract view.

1.1 The utility of computational methods

Computational approaches have provided novel approaches to linguistic problems, and new modeling techniques allow for new insight to familiar problems. Outside of linguistics, physicist Ursula Franklin analyzed the ever-changing role of *technology* on how problems are formulated and addressed: "One has to keep in mind how much the technology of doing something defines the activity itself, and, by doing so, precludes the emergence of other ways of doing ‘it’, whatever ‘it’ might be" (Franklin, 1999). More recently, this is highlighted in artificial general intelligence literature as a need for compositionality, developing different knowledge bases that are able to interact for a higher level representation that may not be modeled in training (Swan et al., 2022).

This maps neatly to linguistics, where a sparser representation represented by minimalist approaches certainly can limit results. Instead, recognizing the many levels of representation - and by extension, what information is needed at each level - can result in deeper understandings of how these levels operate in isolation and in tandem with one another. As such, it’s vital to tease apart these levels and determine how much information is needed for language to be learned.

Even when one focuses specifically on phonological processes, there are still numerous levels at which learning occurs, though these levels nonetheless interact with one another. The most straightforward is learning a phonological rule or constraint, wherein some sequence of sounds triggers a change between the initial representation of the word and how it is produced. These morphophonological changes - processes where a phonological rule is activated because of a morphological alternation - can cue speakers into information about the underlying, abstract representation of the word. For example, take the English suffix /ɪn/, as in ⟨inaccessible⟩ or ⟨indirect⟩. A naive onlooker may assume that this prefix is invariable, but appending it to stems like /pɔsɪbəl/ results in assimilation between the nasal /n/ and labial /p/, resulting in /ɪmpɔsɪbəl/.

Languages with richer morphologies can elicit additional phonological changes that can only be viewed in alternation. In Polish, for example, a particular vowel alternation depends on the structure of the word, only appearing if the next syllable is closed (Gussman, 2007; Rubach, 2016; Kraska-Szlenk, 2007). Given the nominative noun form [vʲɛɕ] (‘countryside’) and the locative suffix /i/, one may believe the locative form would be [vʲɛɕi]. Instead, the vowel is removed, resulting in [fɕi]. In fact, there is a second change in the locative form, where /ɕ/

additionally triggers devoicing in /v/ to assimilate.

These alternations inform numerous other areas of a speaker's phonology, limiting their exposure such that, when speakers build a representation of their language, they can only learn from what is observed. Making generalizations about a language that includes something a speaker has *not* encountered can be difficult. For example, phonotactics describe limitations on sound order. Polish, as we will see, can be comparatively lenient in its phonotactics by allowing clusters that, to English speakers, feel foreign and unacceptable. Halle (1978), for example, uses 'ptak' and 'mgla' as examples of words that English speakers would not accept as English words - but 'ptak' is, in fact, the Polish word for 'bird', and 'mgla' is Russian for 'fog'.[1] English speakers may have words with [pt] internally, such as [æptɪtud], but other than some forced pronunciations of ⟨pterodactyl⟩, [pt] does not occur word-initially. But in language where this cluster *does* occur word-initially, such as Polish, speakers find it much easier to incorporate into their grammars and accept it in new words.

The domains of these questions differ, with alternations looking at an input form and an output form and phonotactics considering the set of acceptable forms in a grammar. However, there is nonetheless some overlap in these problems. This calls into question, do these various levels of phonological abstraction need total access to a fully detailed representation of a word, or are different parts accessed at different points during learning and processing?

While this may be difficult to examine in practice, computational and statistical modeling affords us the opportunity to examine processes in isolation and with sufficient control. To look at models of phonotactic constraints from computational perspectives, linguists make decisions about how much (and what kind of) information is available to the model. From these representations, the model must make some generalization. For example, one hypothesis regarding phonotactics is that more frequent structures are more acceptable. Indeed, more frequent, predictable structures can exhibit reduction (Jurafsky et al., 2001), which can add to variability in pronunciation. Hayes and Wilson (2008) introduces an algorithm that uses

[1] The Polish equivalent, 'mgła', is pronounced [mgwa], which is likely similarly unacceptable to English speakers, with the added bonus of an un-English orthographic character if presented in text. In translating the Russian Cyrillic into the English Latin alphabet equivalent, Halle uses more familiar and interpretable orthography for an English speaking audience to make his point. Though the pronunciation technically diverges, I believe the point holds with Polish.

frequency of feature combinations to generate a weighted list of phonotactic constraints.

This phonotactic knowledge may result from competing underlying information that speakers internalize (Gorman, 2013), rather than dependent on associated frequencies. Durvasula (2020) found that using the algorithm presented in Hayes and Wilson (2008) on English word onset clusters, but making all frequencies equal, produced comparable results when the algorithm was given detailed frequency information. There has been work suggesting that the effect of frequency was variable depending on methods, but still showing improvement over categorical data (Jarosz, 2017; Daland et al., 2011). New categorical algorithms for acquiring phonotactic constraints (Chandlee et al., 2019; Rawski, 2021) present opportunities to probe the question further, determining how well probablistic phonotactic models perform compared to categorical ones.

The level of detail needed for phonotactic learning is also an open question. Speakers may be able to generalize a phonotactic constraint to the relevant features, abstracting away features that do not affect the change at hand (Chomsky and Halle, 1968). These rules find commonalities between many similar processes by creating natural classes. The ability to generalize in this way can aid in learning, reducing the number of cues a listener must attend to when learning some alternation or boundary (Daland and Pierrehumbert, 2011). Models like this can be robust to variation between speakers by requiring amendments to smaller representations, by attending to fewer specifications (Daland and Pierrehumbert, 2011; Kraljic and Samuel, 2007).

Compared to featural representations, a segmental representation is inherently richer, as each symbol represents not only the full featural signature of that phone but also, for segments of two phones or more, constraints on combinations that result in some phonological change. For example, Polish assimilation cannot only be represented by banning adjacent feature bundles with mismatched voicing, as this would ban unvoiced consonants alongside vowels or sonorants. Banning the segment /vɕ/ would be less general, but would allow the segment /ɕi/. Feature representations could allow more specified features to counteract this, but a language may have exceptional segments that require full specification. The Polish word [ssak] ('mammal', nom.) allows a word-initial geminate, which otherwise is very rare in the language. What representation could be most easily adapted to learn to distinguish the segment /ss/ from other, unwanted initial geminates?

Morphological and morphophonological learning has its own set of considerations. With the neural network revolution, morphological learning has embraced many of these frameworks (Kodner, 2022; Kirov and Cotterell, 2019; Rawski and Heinz, 2019). Using general approaches to more specific task can cause unexpected consequences (Gorman et al., 2019), and with neural networks being notoriously opaque and difficult to interpret, it can be incredibly unclear what determined a particular output. Though explainability techniques are an active area of machine learning (Balkir et al., 2022; Yeh et al., 2022; Danilevsky et al., 2020; Doshi-Velez and Kim, 2017), many of these may not apply to the models utilized by linguists, so there can be difficulties in determining what, precisely, was learned. Instead, narrowing the domain of a machine learning problem based on specific linguistic theory can better evaluate how well these models are able to acquire specific patterns before scaling to larger linguistic questions.

Take, for example, the Polish vowel alternation mentioned previously, affecting the word [vʲɛɕ]. This alternation occurs because the main vowel that surfaces is known as a *yer* (Gussman, 2007; Kraska-Szlenk, 2007). Yers result from a historical super-short vowel and are only verbalized when particular criteria are met. They are assumed to be present in the underlying representation of a word and alternating with ∅ (meaning, not being pronounced), rather than being an epenthesized vowel, due to their historical status. There are theories for how yers surface in these alternations (Gouskova and Becker, 2013; Rubach, 2016), but less work done in terms of *learning* yers. When I discuss learning or modeling yers, I specifically mean learning the environments where a yer alternation could potentially occur, as well as where this potential yer would surface in speech or be deleted (as opposed to learning historical patterns or phonetic details). This presents an interesting computational linguistic question - what information must be given to a machine learning algorithm to effectively model how yers alternate? In particular, given that yers make up such a small part of the Polish grammar, what will model yer alternations to the exclusion of the rest of Polish lexicon? Kraska-Szlenk (2007) statistically enumerates the environments in which yers occur, but it is a separate question to see if machine learning techniques are able to replicate these environments.

Questions of phonotactics and morphophonological environments address Sapir's second point of variation, which limits how much an abstraction can vary before it is no longer representative of a language. However, phonological rules must of course have a certain degree of breadth to account for variation among speakers, speech context, and more (Pierrehumbert, 2001, 1994). Constraints must be specific enough to not overgeneralize, but not so specific that common speaker variation can cause difficulties in communication. Here, computational studies may be less effective in modeling what the internal representations of speech sounds look like, but phonetic and psycholinguistic studies adopt similar principles in probing specific parts of a representation. The mechanism that attends to phonetic differences seems closely related to that which acclimatizes listeners to speakers (Mullenix and Pisoni, 1990), but one may ask, how completely do they overlap? Perceptual learning studies show that perceived boundaries between phones can be shifted, but that some boundaries may generalize between speakers only when there are appropriate phonetic conditions that make the change more salient (Kraljic and Samuel, 2007). Other studies examining word access have shown there is some phonological component, since priming a listener on real words can cause difficulty rejecting nonce words (Sumner and Samuel, 2007). If speakers are attending to phonetic features when initially hearing a word, how much of that phonetic representation is stored for recent word access? Examining the effect of phonetic variability on word access will determine whether the form being accessed is dependent on the original speaker's specific vocal features, or if the listener has abstracted away and instead stored the recent access as a representation of the sound absent of phonetic detail.

It is also worth mentioning that many computational linguistic experiments, which address how to best model particular phenomena, often isolate the features of interest in artificial language experiments. In these experiments, researchers can focus on particular factors and remove anything that may not affect the current problem. This approach is important for gaining insight to specific learning mechanisms, but avoids the question of how these processes interact. A phonotactic model may address the specific sound patterns in a language, but if high-frequency items shorten and reduce because of their high predictability (Gahl, 2008; Jurafsky et al., 2001), does the original model represent the new surface forms? After explaining a process in an artificial setting, it is important to ground that in a more natural speech to see how the environment affects the speech act and how representations are stored.

1.2 Contributions of this dissertation

First, I examine factors that affect learning phonotactic representations. To address whether frequency affected phonotactic learning, I look at Polish word-onset clusters to see if statistical models that incorporate word frequency information are better predictors of acceptability than those that use categorical information. I draw from Durvasula (2020), where English word-onset clusters are equally learnable through a gradient, frequency-informed approach as through a simple categorical approach. I also look at the interaction of frequency information with learning window size on acceptability, to determine if the effect of frequency changes with the granularity of the constraint. In chapter 2, I use Polish because of its particularly complex clusters (Gussman, 2007; Jarosz, 2017), due to high tolerance of sonority 'plateaus' as well as asonorous clusters resulting from yer deletion. After conducting an acceptability survey with native speakers, I look at learnability at the cluster, bigram, and featural level, with and without frequency information. I show, using statistical correlations, that as in Durvasula (2020), frequency information does *not* improve prediction of acceptability. I also show that models with finer-grained representations - feature-based constraints or bigrams, compared to a full cluster - also performed worse in these statistical measures.

Then, in chapter 3, I look at how two different machine learning techniques acquire an infrequent morphophonological pattern. I examine Polish yers, which exhibit an atypical alternation pattern unlike traditional insertion or deletion (Kraska-Szlenk, 2007; Gussman, 2007; Gouskova and Becker, 2013; Rubach, 2016). In running these experiments, I evaluate the amount of phonological information needed for these techniques to distinguish words that have yer alternations from those that do not and where they may occur. This interrogates what must be available to formulate a representation of a pattern, as well as what else in the larger lexicon may contest learning. Using a corpus of over 91,000 Polish words (Kirov et al., 2018; Saloni et al., 2015; Woliński and Kieraś; Calzolari et al., 2016), I group words by lemma to find those that have a yer alternation, which make up fewer than 5000 items. Using both a classification and regression tree and perceptron neural networks, I evaluate how these two techniques attempted to learn this small subset of items to the exclusion of the rest of the lexicon, using phonological features and morphological information. I find that particular parameter settings of the perceptron models outperform decision trees, though both learning techniques utilized similar detailed featural and paradigmatic information. This is in contrast to the findings of the phonotactic experiments, where sparser cluster-based representations performed better than ones with featural details. All the code and datasets used in chapters 2 and 3 of this dissertation are hosted on https://github.com/kkostyszyn/DissertationData.

I then step away from a strictly computational focus to address psycholinguistic questions and speculate on its implications for phonological representations. In chapter 4, I conduct an experiment on native English speakers that investigates how voice features affects recognition of nonce words. This will help gain insight to how word access functions in the mind. By extending an established experiment to test multiple voices, I determine whether the phonetic features of each voice are used to facilitate word access or a representation that abstracts away from specific voices is stored in short term. These experiments primed native English speakers on a combination of real words and nonce words before testing their reaction time in recognizing a second, similar word as a real or nonce word. I found no effect of the priming voice, which suggests that, while the voice features may affect the initial perception (Kraljic and Samuel, 2007), the mental representation of a recent word abstracts away from phonetic features. Instead, it calls into question how a listener accepts certain phonetic and phonological patterns into their representation of a word.

After these three experiments, I will summarize my findings and connect their implications to future work in machine learning and computational linguistics before concluding. I find a disparity between the phonotactic and morphophonological experiments, where the first does not benefit much from using a featural representation, while the second finds significant improvement. The results of the psycholinguistic experiment work along those findings to suggest that there are multiple levels of abstraction, and various sound representation processes function at different levels. The phonotactic experiment put more weight in the full cluster representation in a similar manner to psycholinguistic word access results, which supported a level of word access absent of specific phonetic details. Meanwhile, the morphophonlogical results support incorporating phonological features, which recalls how salient phonetic information can affect perceptual learning.

2: PHONOTATICS

2.1 Introduction

Polish is known for complex onsets that defy principles of sonority sequencing (Gussman, 2007; Jarosz, 2017; Zydorowicz and Orzechowska, 2017), which are of significant interest for acquisition models. These complex onsets are a prominent feature among Slavic languages, including Slovak (Bárkányi, 2011). In Polish, these are exemplified by common interrogatives such as [ktɔ] ('who') and [gdɪ] ('when, if'), where the onset cluster of two plosives constitute a sonority plateau (Jarosz, 2017; Jarosz and Rysling, 2016), where sonority neither rises nor falls. There are also instances where sonority will fall in a way contradictory to the SSP, as in [mʃa] ('mass', nom.), as well as instances where within a single cluster, sonority will both rise and fall (or fall and rise). These can be seen word-medially in [ja.bwkɔ] ('apple', nom.), or word initially in and [mgwa] ('fog', nom.). In a count of sonority profiles in child-directed speech (Jarosz, 2017; Haman et al., 2011; Jarosz et al., 2013; Weist et al., 1984; Weist and Witkowska-Stadnik, 1986), tokens that show a sonority fall make up less than 1% of total tokens.

Because of the complex phonotactics at play in Polish clusters - in particular, elaborate defiance of sonority sequencing - they are a testing ground for studying the ability of different phonotactic models to predict phonotactic acceptability. Previous work on Polish onsets has examined the interaction of phonological rules and the implications of gradience in a lexicon in a specific model, particularly the idea that lexical gaps are often a result of rules failing to interact (Gorman, 2013). This, of course, would push against phonotactic models based on probability (Hayes and Wilson, 2008) and other gradient measures, because phonological rule application is a categorical process. It has also been argued that categorical models sufficiently describe the same phenomena as gradient models (Durvasula, 2020), which is a departure from studies that focus on the effect of frequency (Hayes, 2011; Jarosz et al., 2013; Jarosz, 2010). I examine the hypothesis that categorical grammars perform as well as, if not better than, gradient grammars in predicting acceptability of word-onset clusters in Polish.

I surveyed a group of native Polish speakers for their judgments on nonce words made with Polish-like onset clusters. I trained two sets of three Polish phonotactics models. One set uses frequency to inform regressions, while the other set uses categorical information, based on whether the onset appears in a Polish dictionary as a 'familiar' word. Each set is made up of a cluster model, a bigram model, and a feature-based model. The cluster and bigram models directly measure correlations between scores from the acceptability survey and either the frequency or categorical familiarity score. The categorical feature model uses an algorithm that generates a phonotactic grammar based on a partially ordered structure built from constraints that are never violated (Chandlee et al., 2019; Rawski, 2021). This algorithm returns a list of inviolable constraints represented as feature matrices. The gradient feature model uses a maximum entropy learner (Hayes and Wilson, 2008), which returns similar constraints, as well as weights for each constraint depending on how violable they are given the input data. Given these six models, I compared the resulting correlations between acceptability scores and familiarity/grammaticality. The results show that the categorical models generally outperformed the gradient model.

In the first section of this chapter, I discuss work on phonotactics that examine both gradient and featural models. I then lay out previous research on Polish phonotactics and onset clusters and report results from an acceptability survey that determines the acceptability of attested Polish clusters in novel words. I then present the results of the categorical models first, then the gradient models, moving in each section from the most general cluster representation to most specific featural representation. Finally, I compare the results from each model to one another, showing which ones had the highest predictive power and what the failings of each model may be, before concluding.

2.2 Previous work

2.2.1 Phonotactic modeling

Phonotactic models use various parameters and representations for human speech, taking some input and judging it to be acceptable or unacceptable, in a gradient or categorical manner. These parameters are rarely arbitrary, and reflect contextual information which informs the grammatical model. For example, a model must choose whether to refer to phones as symbols, or to refer to their individual features when determining constraints. A feature-based model can make wider generalizations about what combinations of adjacent features are grammatical than a symbol-based approach. A categorical feature-based model would then determine at which point adding a certain feature would render a representation ungrammatical (Chandlee et al., 2019), while a probabilistic model would determine the likelihood of these features being grammatical (Hayes and Wilson, 2008).

Then, the model must specify what which sequences of phones it will consider. If a model will examine any phone in adjacent proximity, up to some window of size *k*, it's considered *strictly local* (Rogers and Pullum, 2011), but if it is instead limited to *k*-sized window of phone after they have been projected to a tier for examination, it is instead *tier-based strictly local* (Heinz et al., 2011). Alternatively, if the model is concerned only with the order of the phones, rather than their adjacency, then the model is considered *strictly piecewise* (Heinz, 2010; Rogers et al., 2010). Depending on these choices, the evaluation may return a completely different analysis. Processes like vowel harmony are not suited to strictly local accounts, for example, since any intervening consonants would force the window size *k* to increase arbitrarily, without meaningfully adding to the analysis.

Another parameter is whether a model is categorical or gradient. As an example of the former, Chandlee et al. (2019) use partially-ordered representations to determine which combinations of features are grammatical. As an example of the latter, Hayes and Wilson (2008) use the principle of Maximum Entropy to return what combinations of features are likely to be grammatical. The latter may be more forgiving to variation between speakers, and even within a single speaker, but may also require additional data and processing to determine these probabilities. A categorical model, in comparison, treats the existence of a structure in the data as sufficient evidence for acceptability.

2.2.2 Gradient vs. Categorical Models

Many have argued that phonotactics models should account for gradient judgments. Hayes and Wilson (2008, p. 382) advocate for this position, writing,

> "All areas of generative grammar that address well-formedness are faced with

> the problem of accounting for gradient intuitions... In the particular domain of phonotactics, gradient intuitions are pervasive: they have been found in every experiment that allowed participants to rate forms on a scale... Thus, we consider the ability to model gradient intuitions to be an important criterion for evaluating phonotactic models."

Their probablistic model is based on the concept of maximum entropy. The crux of this approach is that "well-formedness can be interpreted as probability," and that the "probabilities assigned to forms by a maxent grammar will correspond to the well-formedness judgments of native speakers, with lower probabilities for forms judged less acceptable" (Hayes and Wilson, 2008, p. 383). Here, a *maxent* grammar - short for a *maximum entropy* grammar - assigns weighted constraints as probabilities, where a higher weight on a phontactic constraint means it is less likely to be violable. Because gradience surfaces in native speaker judgements of rare forms, Hayes and Wilson (2008) argue that a phonotactic model *must* model gradient intuitions. Because of the success of this seminal work, this approach has become standard for modeling phonotactic judgments.

Rather than a gradient grammar, which allows for varying levels of acceptability based on salient features, a categorical grammar will have binary membership determined by some salient parameters. A relevant example from Buckley (2001) concerns a vowel-raising rule in Polish where, given a word that has /ɔ/ in the final syllable *and* that syllable has a voiced coda, the /ɔ/ raises to [u] (Kraska-Szlenk, 2007). This can be seen in the difference between [mɔ.ja] ('my', fem. nom. sing.) and [muj] ('my', masc. nom. sing.), where the feminine suffix /-a/ blocks raising. Buckley argues that this vowel-raising rule persists in novel words, not by its own productivity, but because novel words may express a similarity to words known by the speaker that exhibit this vowel-raising. Here, the categories of interest were 'vowel-raising in inflection' and 'no vowel raising in inflection', and membership was decided based on similarity to pre-existing members of the group. Novel words that contained the appropriate environment to trigger raising, but did not match an existing word with raising, were in the 'no vowel-raising' category, and thus belonged to a different inflectional paradigm. This is followed by more recent work on how irregular morphological patterns are acquired (Yang, 2016), which suggest that the infrequency of the pattern blocks speakers from acquiring it as a regular rule.

Gradience, of course, occurs in the results, but rather than being a requisite part of the model, it can instead be a result of categorical interaction. For example, lexical gaps are an area of the lexicon where a word, based on a language's phonology, *should* exist, but do not. However, these gaps may instead be a result of categorical interaction (Gorman, 2013). The interaction between two phonological rules may systematically eliminate the case in which an otherwise licit word may occur, without relying on frequency-based statistics.

In Durvasula (2020), comparative studies supports this hypothesis for two data sets of accceptabilty judgments on English onset clusters (Scholes, 1966; Albright, 2007). In one study, Durvasula compared the learner in (Hayes and Wilson, 2008) (henceforth, HWPL for 'Hayes-Wilson Phonotactic Learner') with a simple, categorical phonotactic learner. The grammar for the categorical leaner contained all the observed single feature unigrams and bigrams, as well as all the observed *segment* unigrams and bigrams in the learning corpus (Durvasula, 2020, slide 36). The evaluations showed that this categorical model was at least as good as the HWPL in accounting for the reported gradient judgments.

Durvasula also compared two instances of HWPL: one trained with with actual type frequencies, and the other trained with equalized frequencies (to neutralize the information frequency provides). Durvasula showed that the grammar generated by the learner with frequencies equalized was able to make similar predictions to that of the model with the actual frequencies, concluding that distinctive frequency information was irrelevant. In fact, he argues that added gradience can harm the performance of the model by showing that, in some cases, a model learning weighted constraints had weaker predictions than one learning categorical constraints.

In sum, these results cast doubt on the claim that gradient phonotactic learning models are necessary to account for gradient acceptability judgements and are inherently better than categorical learning models. In this chapter, I extend Durvasula's findings by examining acceptability of Polish onset clusters. I previously had conducted three of the six studies (Kostyszyn and Heinz, 2022), shown in figure 2.1. Here, I expand the study a full 2x3 design. It not only compares categorical and gradient models, but also the level of information being

CATEGORICAL	GRADIENT
Cluster	Clusters with type frequency
2-local categorical	Probabilistic bigram model
BUFIA	**HWPL**

Figure 2.1: Configurations of categorical vs. gradient experiments.

accessed - the full cluster, bigrams derived from the cluster, or the featural combinations that make up the clusters.

2.3 Polish clusters

First, I will describe how the Polish cluster data were collected. After that, I will discuss the acceptability survey created using these clusters and give an overview of the results, which are used to evaluate the statistical models.

2.3.1 Polish data

To identify the acceptability of sound sequences in word-onset clusters in Polish, I extracted onsets from three main corpora and assessed their wellformedness with native speakers. The corpora described here also informed the training data used for the modeling experiments.

The most sizable corpus used was the Web 1T 5-gram corpus (Brants and Franz, 2009) (henceforth known as Web 1T), a text corpus taken from a forum of Polish speakers on the internet. All word-initial consonant clusters were extracted from this corpus. However, due to this being an internet forum, a variety of irrelevant material occurred in this output: typos, foreign characters such as hiragana, and non-linguistic material such as hyperlinks. To differentiate between 'legitimate' and 'illegitimate' clusters, each cluster obtained from Web 1T was retained only if it was also found in at least one of two other sources. One was the Polish data from the Universal Morphology project (Kirov et al., 2018) (henceforth known as Unimorph), using text adapted from a inflectional dictionary specific to Polish (Saloni et al., 2015; Woliński and Kieraś; Calzolari et al., 2016). The other was a list of bigrams counted in corpus of child directed speech, phonetically transcribed from speech (Jarosz, 2017; Jarosz et al., 2013; Weist et al., 1984; Weist and Witkowska-Stadnik, 1986) (henceforth known as the Weist-Jarosz list).[1] This resulted in a

[1] The Weist-Jarosz list was written in IPA originally, but using slightly different conventions, namely the

list of 329 unique clusters, with frequencies based on the number of occurrences in Web 1T.

Because Polish orthography maps very closely to its phonology, extraction and transcription of the clusters from Web 1T was done at the same time by use of a finite state automaton built with Pynini (Gorman, 2016). Only words that began with a consonant were accepted and rewritten to their IPA counterpart; all vowels, any character following a vowel, and any invalid character (such as the aforementioned hiragana, but also including numbers) were rewritten to the empty string to remove them.[2]

It is worth noting that two clusters that existed in the corpora were removed at this point because of the likelihood that speakers would misread the surface form. The written clusters ⟨dż⟩ and ⟨cz⟩ most commonly map to the affricates /d͡ʒ/ and /t͡ʃ/, respectively. However, the generated onset list also included the non-affricated /dʒ/ and the cluster /t͡sz/, an affricate followed by a sibilant. These onsets would also be written as ⟨dż⟩ and ⟨cz⟩, respectively. Because the affricates /d͡ʒ/ and /t͡ʃ/ generally occur with a much higher frequency, they are therefore more likely to be how these characters would be read by Polish speakers. Because the acceptability survey below was conducted through written text rather than recorded voice, the clusters /dʒ/ and /t͡sz/ could not be adequately judged for acceptability and so were removed from the data.

2.3.2 Acceptability survey

A survey was conducted with the 329 clusters described in 2.3.1. The list of clusters was written in IPA, appended to the nonce stems ⟨arek⟩, ⟨olek⟩, and ⟨ąc⟩. These stems were chosen to avoid conflict with ⟨i⟩ which can signal palatalization, as well as to allow interaction with nasal vowel ⟨ą⟩. This resulted in a list of 987 words. The nonce words were then reverse transcribed back into Polish orthography and presented to the speakers. Four native speakers were asked to give binary acceptability judgments for all nonce words, where '0' denotes unacceptable, and '1' denotes acceptable. Stimuli were presented in text form, randomly ordered. Since there were four speakers and three words for each cluster, each cluster received 12 ratings. These 12 ratings were converted to a percentage, which can be seen in appendix A.1 for all clusters. Figure 2.2

distinction between [ʃ] and [s]. Corresponding symbols were adjusted for consistency.

[2] One exception to this was the case of palatalization, which is denoted in Polish orthography with the vowel ⟨i⟩, where the sequence ⟨si⟩ would map to /ç/ rather than /si/. In this case, the relevant substring is mapped to the palatalized consonant, rather than the empty string.

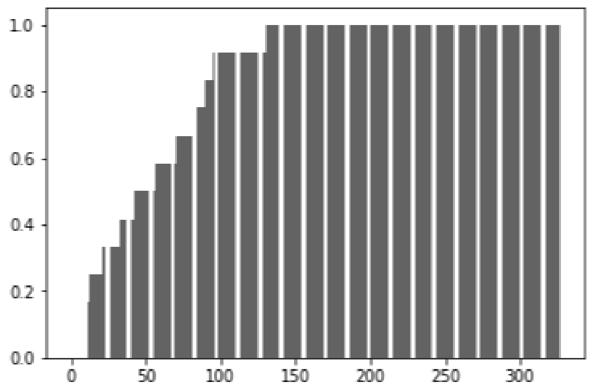

Figure 2.2: Distribution of results from acceptability survey. Each cluster received 12 judgments (given three possible words for four speakers), which is represented as a decimal along the Y-axis here.

shows the distribution of acceptability scores in the surveys, to emphasize that the categorical decision nonetheless resulted in gradient results. These ratings are provided in a file in the repository at https://github.com/kkostyszyn/DissertationData.

2.4 Experimental design

I investigated six configurations of categorical vs. gradient data, combined with either a segmental, bigram, or featural level of detail. These configurations can be seen in figure 2.1. For all models, I use statistic regressions using Pearson's r (Pearson, 1895) and Kendall's τ (Kendall, 1938). These were chosen to examine the strength of the relationship of cluster acceptability (as determined by survey) and the corresponding frequency or wellformedness score, described for each model in their dedicated section below. Pearson's r is intended for linear relationships, while Kendall's τ is intended for rankings - because they take different approaches, these two statistical methods can give more insight to what arrangement of data predicts acceptability best.

For the Bottom Up Factor Inference Algorithm (Chandlee et al., 2019; Rawski, 2021) and the Hayes-Wilson Phonotactic Learner, each algorithm outputs the constraints found. I additionally examine generalizations in each set of outputs to better inform how to interpret the statistical analyses. In particular, I evaluate the constraint lists on the set of test clusters to determine which algorithm better covers the data and results in fewer violations.

2.5 Categorical models

The categorical models decided a possible onset was well-formed if it occurred in at least two words in the Słownik Języka Polskiego ('Dictionary of the Polish Language') hosted by Wielka Encyklopedia Powszechna PWN (henceforth known as PWN). PWN is a longstanding Polish institution, making it trustworthy for this purpose. The general assumption was that these clusters would approximate the ones in the mental lexicon of Polish speakers. Clusters that occurred exactly once were excluded on the grounds that those words were rare.[3]

Of the clusters in the acceptability survey, they were marked with a 1 if the categorical model found them ill-formed and 0 if well-formed. This results in negative correlations between speaker-judged acceptability and binary well-formedness for the full cluster model and 2-local bigram models. The feature-based model, conducted later, used an evaluation system that reverses these scores. So, while values will be reported here as calculated, the absolute values of these scores will be used for discussion later.

2.5.1 Full cluster

The full cluster model determined the predictive relationship between a 'familiar' cluster, one present in PWN, and the acceptability score of the clusters in Web1T. The correlation between the cluster model's scores and the average speaker judgments for the 329 clusters, as measured by the Pearson's correlation coefficient, was $r = -0.755$ ($p < 2.2e - 16$). Thus, we can see that there is a strong and significant correlation between familiarity with at least two lexical items beginning with a certain onset cluster, and acceptance of other words that have the same onset cluster, regardless of violations of sonority. Using Kendall's rank coefficient, the correlation is $\tau = -0.7238829$ ($p < 2.2e - 16$). Again, the correlation is high enough to justify stating that there is a strong relationship between onset familiarity and acceptability.

2.5.2 2-local bigrams

Using the cluster list as a base, each cluster was augmented with # as start and end boundary symbols, then divided into its bigrams. The bigram models divided the clusters found in PWN

[3] In future work, it may be preferable to select words that meet a more sophisticated frequency threshold.

into bigrams to create 'familiar' bigrams. The acceptability score of a word is equal to the number of illicit bigrams in its onset clusters. So again a perfectly acceptable cluster would score '0'. Consequently, all words scored '0' by the cluster model will be scored '0' by the bigram-type model as well because if a cluster is attested in the training data, all of its internal bigrams are attested in the training data as well. Conversely, there are potentially clusters that could be scored '0' in the bigram model because its bigrams appear individually in PWN, even though the full cluster does not (and so the cluster model would score such clusters '1').

The correlation between the Strictly 2-Local model's scores and the average speaker judgments for the 329 clusters, as measured by the Pearson's correlation coefficient, was $r = -0.726$ ($p < 2.2e - 16$). Kendall's rank correlation was $\tau = -0.693$ ($p < 2.2e - 16$). Again, high values suggest that having seen a bigram is a good predictor of whether speakers would accept that bigram in a novel word onset.

2.5.3 Learning with features

The Bottom-Up Factor Inference Algorithm or BUFIA for short, is a categorical algorithm that uses partially ordered representations to determine what combinations of features are possible in a sequence of phones (Chandlee et al., 2019). These sequence extend up to length k and have n features per factor. For example, the possible constraint *[+A, +B, +C] would have $k = 1, n = 3$, while the constraint *[+A][+B][+C] would $k = 3, n = 1$. So, a constraint *[+A] entails constraints *[+A, +B] and *[+A][+B]. Constraints with larger values of k would be able to track longer distance relationships, while larger values of n can find finer interactions between features. This additional expressivity comes at a cost, since the runtime of BUFIA is exponential. Increasing either k or n by 1 increases the number of possible permutations considered.

Given a list of word forms and the feature chart for a given language, BUFIA considers all featural combinations that surface for the given values of k and n. If the word does *not* exhibit some combination, it is added as a possible constraint for consideration; for all models below, I used principle 1 for gauging possible constraints: "A constraint is only added to the grammar if its extension is not subsumed by the extension of the current grammar" (Rawski, 2021). This

	PEARSON	p-value	KENDALL	p-value
C_K2N2	-0.033	0.5475	-0.034	0.5193
H_50	0.031	0.5696	0.0432	0.4135

Table 2.1: Pearson and Kendall correlation values for BUFIA.

process outputs the list of final found constraints, in order of discovery. For evaluation, words are assigned the number of the discovered constraints which they violated.

I used two implementations of BUFIA, one written in Haskell by Professor Jeffrey Heinz [4] and one written in C++ by Logan Swanson[5]- both are referred to by the implementation language for convenience. The Haskell version included options to focus on subsets of features, while the C++ implementation ran faster and printed out constraints as they were found rather than at the end, which allowed for more flexibility in the face of memory limitations. I ran seven model configurations to compare the resulting featural combinations. These were as follows, marked with (H) for the Haskell implementation and (C) for the C++ implementation:

1. the first 50 constraints when $k = 3, n = 3$ (H) (listed as H_50 in 2.1)
2. all constraints before memory ran out, when $k = 3, n = 3$ (C)
3. all constraints when $k = 2, n = 2$ (H)(C) (listed as C_K2N2 in 2.1)
4. manner features, with $k = 3, n = 3$ (H)
5. manner features, with $k = 2, n = 2$ (H)
6. place features, with $k = 3, n = 3$ (H)
7. place features, with $k = 2, n = 2$ (H)

The first two models are the broadest in that they do not limit to any subset of the feature chart. Instead, they compare the search space of the algorithm based on the factor and feature limits. The first model, run in the Haskell implementation, had larger values of k and n to allow for more precise constraints to be found. However, the permutations resulting from these larger values would also have considered many more combinations of features, which exceeded the

[4] https://github.com/heinz-jeffrey/bufia

[5] https://github.com/pterodactylogan/bufia

capacities of the laptop on which these experiments were run. So, the list was limited to the first 50 found.

In evaluation, this model performed with a Pearson's correlation coefficient of $r = -0.03328173$ ($p = 0.5475$) and Kendall's rank correlation of $\tau = -0.03395568(p = 0.5193)$. Neither value is significant, and any predicted trend would also be extremely small. While this appears at first glance to imply that these implementations of BUFIA perform dismally compared to the models already discussed, examining the resulting list of constraint violations suggests a more complicated story. Of the 329 clusters used to test each model, 316 clusters had 0 violations given the the first list of constraints. So, it may instead be the case that the constraints output by this model covered the data so well that there were too few violations from which to calculate a strong correlation.

The second model, run in the C++ implementation, had similarly large factor and feature limits. However, because constraints were printed as they were found, it was possible to run the algorithm until memory ran out, meaning that the main limitation was imposed by the laptop. Nonetheless, 362[6] constraints were found before crashing.[7] This time, the Pearson correlation was $r = -0.06925336$ ($p = 0.2103$), and the Kendall rank correlation was $\tau = -0.0177016$ ($p = 0.7363$). Again, the correlations are very small and nonsignificant. Notably, despite this iteration having over 300 more constraints than the first model, the exact same number of test items had constraint violations. In this iteration, certain clusters had *more* violations than the first model. For example, the cluster [pʃtʃ] had 6 violations when evaluated with 50 constraints and 22 violations when evaluated with 362.

On the other hand, the broad $k = 2, n = 2$ model was run in both implementations of BUFIA to compare the order in which constraints were found. The smaller search space meant that the algorithm successfully terminated for both the Haskell and C++ implementations, allowing for most direct comparison of the order in which constraints were found. The Haskell implementation

[6] Logan Swanson later ran the same configuration on Stony Brook's SeaWulf cluster to see how many additional clusters could be discovered if BUFIA completed without crashing. Despite the extra hours and memory, he found 363 - only one additional constraint.

[7] This iteration ran for just under one hour before crashing. To compare, when generating $k = 3, n = 3$ constraints for Haskell *without* limiting to the first 50 constraints, it took 8 hours before crashing. However, the Haskell implementation would not print constraints without completing the constraint generation process, so these results could not be used.

found 121 constraints, and the C++ one found 118. The difference is due to the different ways that the Haskell and C++ implementations of BUFIA search the constraint space. The order in which constraints were considered meant that banned factors were accounted for by potentially different constraints. The Haskell implementation ordered constraints by length k, while the C++ implementation ordered constraints by measuring the number of differences from the unique 'empty' constraint *[]. For example, the first five constraints of both lists are as follows:

- *[–Anterior,–Distributed]
- *[–Consonantal][+Round]
- *[–Consonantal][+Lateral]
- *[–Consonantal][+Trill]
- *[–Consonantal][+LabioDental]

There is only one constraint of length 1, [–Anterior,–Distributed]. It is the first constraint for Haskell, because all constraints of length 1 are found before constraints of length 2. For C++, it is the first in terms of presentation, but it has a distance of 2 from the empty constraint . The next four constraints have $k = 2$ and distance $d = 3$. However, the next constraint for the Haskell implementation is [-Consonantal][+Sonorant,-Labial], which still has a single feature in the first factor but has two in the second. Since C++ is organized by distance, the next factor is [+Round][-Consonantal]. This same constraint is found by Haskell, but much later in the runtime, as the 21st constraint discovered.

For the Haskell implementation, the Pearson correlation was $r = -0.033$ ($p = 0.5475$), and the Kendall rank correlation was $\tau = -0.034$ ($p = 0.5193$). For the C++ implementation, the Pearson correlation was $r = 0.002$ ($p = 0.9719$), and the Kendall rank correlation was $\tau = -0.032$ ($p = 0.5413$). Again, no values are significant. As in the first two models, 316 clusters had no violations, and among the remaining 13 clusters *with* violations, the number of violations varied due to the difference in constraint discovery procedures. Looking at the cluster [tʒt͡s], the Haskell implementation had two fewer violations than the C++ implementation (2 compared to 4), despite finding more constraints (121 to 118).

After these models, I looked at four that considered a select subset of features. The 'manner' model looked at the features Consonant and Sonorant, while the 'place features' model look at Coronal, Labial, and Dorsal. Each of these were first run with settings $k = 2, n = 2$. The manner model resulted in *no* constraints at these settings, as all permutations of features and factors were accounted for in the data. The place model, on the other hand, found three constraints, but when evaluated on the test data set, found no violations.

I ran these models each again with $k = 3, n = 3$ to see if there were any longer distance relationships found. Both models found more constraints, with 7 constraints for the manner model and 33 for the place model. However, evaluation still found zero constraint violations in the test set for both models, resulting again in standard deviations of zero.

2.6 Gradient models

As mentioned, frequency was determined based on number of occurrences in the Web 1T corpus (Brants and Franz, 2009). Recall that the categorical models were marked 1 for a *violation*, 0 otherwise. The *gradient* cluster and bigram models use frequency from Web 1T to estimate how often a speaker would see this cluster, rather than considering violations. This will result in a positive correlation to predict the same relationship instead of a negative one.

The HWPL maximum entropy model, on the other hand, returns constraints based on the frequency of clusters, and a corresponding (positive) weight. The weight is higher for constraints that are less often violated. The resulting statistics thus predict the likelihood that a highly weighted, infrequently violated constraint suggests a highly acceptable cluster, which is a negative correlation.

As before, all results here will be reported as generated, but discussion will be based on the absolute values.

2.6.1 Clusters with type frequency

The cluster model with frequency looks at the correlation between frequency, extracted from Web 1T, and the acceptability scores. Using Pearson's correlation coefficient, $r = 0.121$ ($p = 0.03148$). Using Kendall's rank correlation, $\tau = 0.246$ ($p = 3.288e - 08$). Both of these values are significant, but predict much weaker correlations than the corresponding categorical model.

2.6.2 Probablistic bigrams

Acceptability scores were, again, determined by the number of illicit bigrams. Additionally, frequencies were calculated by adding together the frequencies of each cluster that contains the bigram. For example, consider the bigram ⟨#b⟩. To determine its frequency, we look at all full-size clusters that contain this bigram - in this instance, let us assume the only two clusters are ⟨#br⟩, which occurs 10,501 times and ⟨#bl⟩, which occurs 8,908 times. Given these clusters, the bigram ⟨#b⟩ would have 19,409 occurrences.[8] With this in mind, Pearson's correlation coefficient comes out to $r = -0.122(p = 0.03026)$ and Kendall's τ as $\tau = -0.305(p = 3.476e - 11)$, both of which are significant and, like the cluster models, are predict weaker correlations than their categorical counterparts.

2.6.3 Maximum Entropy

Of the 100 constraints returned by HWPL, 12 were scored with a weight above 6. Constraints with lower weights are 'more' violable, which makes it difficult to generalize how well they represent the input data. The two most highly-weighted constraints referred to features within a single phone: *[–Anterior, –Continuant, +Strident] (referring to affricates such as [t͡ʃ] and [d͡ʒ]) and *[–Anterior, +Strident, –DelayedRelease] (those same affricates, as well as palatalized sibilants). Only one of these top 12 constraints referred to a boundary symbol: *[-Anterior,-Continuant][-word_boundary]. This class largely referred to affricates or palatalized consonants. As noted, a palatal consonant is often written orthographically followed by ⟨i⟩ in this position, and elsewhere in the cluster will be diacriticized.

The other highly-ranked constraints, in order of descending weight, were as follows:

- *[–Continuant, +Strident][–Round, –Voice]
- *[–Voice, –DelayedRelease][–Sonorant, +Anterior, +Voice]

[8] This assumes zero occurrences of ⟨#b#⟩ as an onset cluster as well. As it stands, ⟨#b⟩ as a *bigram* has a real frequency of 107,804

- *[–Dorsal, –Voice, +Continuant][–Anterior, +Voice]
- *[+Labial, +Voice, –Nasal, –LabioDental][+Coronal, –Continuant]
- *[–Round, –Voice, +Continuant][–Sonorant, +Voice, –LabioDental]
- *[–Labial, +Continuant, +DelayedRelease][–Round, +Voice, +Continuant, –Lateral, –Trill, –LabioDental]
- *[-Continuant, +Strident][+Voice, –Strident, –Lateral, –Nasal]
- *[+Sonorant, –Round, –Lateral][+Anterior, +Strident]
- *[+Dorsal, +Voice, +Continuant][-Coronal, +Dorsal, +Continuant]

In three of these constraints, the HWPL learner does note the voicing assimilation within clusters in Polish. Other constraints are more informative if the combination '-Round, -Voice' is simplified to only '-Voice', since [w] was our only round consonant.

First, I looked at the relationship between the weights assigned by the HWPL and the speaker's judgments. Pearson's correlation coefficient was $r = -0.06955772$ ($p = 0.2083$), which was not significant. On the other hand, Kendall's $\tau = -0.14244414$ ($p = 0.004941$) was significant.

Second, I used the evaluation function built into BUFIA using the constraint list found by the HWPL. Rather than comparing speaker judgments against the weights directly, this approach determines whether the constraints adequately cover the clusters found, or if there are phonological violations that the HWPL did not learn. Using this approach, Pearson's $r = -0.1455897(p = 0.008174)$ and Kendall's $\tau = -0.1889753(p = 0.0002421)$, both of which are significant.

2.7 Discussion

The cluster and bigram categorical models demonstrated that, as expected, acceptability would be predicted based on that same cluster occurring elsewhere in the lexicon. This mirrors other research in Polish, not on yers but rather on analogy of a vowel-raising process (Baranowski and Buckley, 2003); in this study, speakers were asked to generate words containing a particular

		Cluster	Bigram	Features	(BUFIA Eval)
Categorical	Pearson	**0.755**	**0.726**	—	0.031
	Kendall	**0.724**	**0.693**	—	0.043
Gradient	Pearson	**0.121**	**0.122**	0.070	**0.146**
	Kendall	**0.246**	**0.305**	**0.142**	**0.189**

Table 2.2: Absolute value of correlations for each model. Significant values are in bold. For ease, I only show the correlation for the H_50 model to represent BUFIA's result, though the scores for any models other than the place and manner models are also suitable.

final-syllable vowel environment, which historically triggered a vowel-raising process. Findings showed that the greatest indicator of raising was similarity of the nonce word to extant words in the lexicon which exhibited raising. This attributes acceptability to analogy across the speaker's grammar, rather than a specific synchronic process.

However, an unexpected result was that the number of illicit bigrams per cluster was a weaker predictor than the binary indication of whether a cluster existed or not, per table 2.2. The difference was slight, but consistent across measurements. The difference may have been more extreme if there were more 4- or 5-gram clusters, containing bigrams that did not occur elsewhere in the lexicon as an onset in isolation. As it stands, it shows that speakers are in fact referencing their own lexicon for previously-established grammatical clusters, but they may not be looking more deeply at the contents of the clusters, for specific features in sequence.

The BUFIA models had very poor statistical results, but as mentioned, in the context of violations within the test set, they performed well in the sense that violations were confined to a select few clusters (and the same clusters across all models). Comparatively, the HWPL learner had significant, though weak, correlations, but many more violations. The BUFIA-generated models (excluding models that focused on particular features) had between 50 and 362 constraints, and in evaluation, each models had 46 clusters with violations. The HWPL had 100 constraints, and 96 clusters with violations - more than twice the violations of the BUFIA models. This suggests that, for Polish, the earliest discovered constraints from the bottom-up factor approach are more likely to cover the phonotactics combinations within the language than a frequency-led approach.

Compared to their corresponding categorical models, the gradient models performed very poorly. For these models, the ranked coefficient correlation performed twice as well as Pearson's correlation for all models. The evaluation function used by BUFIA is the exception, but this approach treats constraints as though they were determined categorically. Nonetheless, the ranked correlation gives a slightly stronger relationship. Because all of the categorical models found stronger relationships using Pearson's correlation, the change in direction supports that Kendall's ranked correlation performs best on gradient information because it is intended for ordinal information (Kendall, 1938), while Pearson is intended for linear relationships (Pearson, 1895).

Regardless of discrepancies between correlation measures, both measures were much lower for gradient models than for categorical ones. This suggests not only that frequency is a worse predictor than mere familiarity, but more surprisingly that frequency has a very weak effect overall, and that language speakers can internalize acceptable clusters from very few examples. As before, the models using bigrams resulted in slightly weaker correlations than those using the full cluster. For Polish specifically, this may be a result of the complexity of the consonant clusters. 'Sonority plateaus' (Jarosz, 2017), where subsequent phones within a consonant cluster have equal sonority, are acceptable in Polish. Moreover, because deletion of a yer in an onset cluster can result in *falling* sonority instead of rising, using bigrams to measure means that rising sonority cannot reliably be used to find cluster boundaries. For example, the cluster [ss] comes from the word [ssak] ('mammal', nom.), and was rated acceptable for all test forms by all speakers. Because ⟨#s⟩ and ⟨s#⟩ are valid bigrams in addition to ⟨ss⟩, this results in sequences like [sss] of unbounded length. Similarly, [tk] and [kt] are both valid bigrams, which could allow for ⟨ktktkt⟩ as a viable cluster. The cluster model can account for this by more strictly limiting options, while the bigram model will need additional constraints to limit potentially infinite sequences or multiple changes in sonority.

A criticism of these results is that a categorical survey, such as the one used here, will be unfairly biased toward categorical results. However, Hayes and Wilson (2008, p.382) write "it is an inherent property of maximum entropy models that they can account for both categorical and gradient phonotactics in a natural way." Additionally, as shown previously in Figure 2.2, judgements are nonetheless gradient between speakers and word stems.

However, it must still be highlighted that these models are predicting the acceptability of

	occurring	non-occurring
well-formed	brick	blick
ill-formed	sphere	bnick

Figure 2.3: Examples of ill- and well-formed words, and occurring and non-occurring words (Hyman, 1975).

attested clusters. Albright (2009) notes that the HWPL learner has difficulty learning differences in acceptability between attested clusters. This is a limitation it shares with BUFIA - because these algorithms are first and foremost considering *attested* clusters, from real Polish language sources, they all must be accounted for in the grammar. Moreover, HWPL is outperformed by a feature-based bigram model and a segment-based bigram (Albright, 2009), again when restricting evaluation to attested clusters. So, algorithms like BUFIA and HWPL may be better suited for determining phonotactic constraints over the grammar entirely, rather than ranking clusters within the grammar against one another.

It may be the case that these representations do not capture the realities of speech closely enough. It had been well studied (Gahl, 2008; Jurafsky et al., 2001) that words that are low-frequency and less predictable have more careful speech, while those that are highly predictable can be reduced. This reduction may account for the inability to discern acceptability between attested clusters. Take, for example, the Polish word [‘ja.bwkɔ] (‘apple’, nom.), with the word-medial cluster /bwk/. Polish prefer complex onsets to syllables with codas (Gussman, 2007), and stress regularly falls on the penultimate syllable (except in loan words), so it’s evident that /bw/ is not a separate syllable. In natural speech, however, /w/ is reduced to deletion, which triggers voicing assimilation, resulting in [‘ja.pkɔ]. To use an example at the word onset position, [ssak] (‘mammal’, nom.) has been discussed in terms of its rare word-initial geminate. Word-medial geminates are common and contrastive - compare [rɔ.‘d͡ʒi.nɪ] (‘families’, nom.) to [rɔ.‘d͡ʒin.nɪ] (‘familial’, adj.) - but this length distinction in word-onset position may be reduced in natural speech, if the word is still highly predictable.

What makes these results interesting in a larger scheme is the exploration of sound sequences that are judged to be ill-formed, but still occurring (Algeo, 1978). A complement to lexical gaps, these words, like ⟨sphere⟩ in Table 2.3 should not play a part in the grammar since they’re already judged to be ill-formed. To my knowledge, there is no term for these occurring but ill-formed words, sometimes called 'structurally excluded' (Gorman, 2013) words, so I refer to these clusters as 'frozen'. These clusters in Polish may continue to play a frozen rule in the lexicon; these infractions may not be able to be generalized into constraints, but rather continue as a single unit.

These 'frozen' clusters contrast with English, where the onset [sf] does not appear to be permissible in novel words. Other words beginning with [f] can be created (or already exist) such as [fir], [fɪn], or [fɪt]; if the onset is changed to [s], the words [sir], [sɪn], or [sɪt] are equally legal and have meaning. However, if the onset is changed to [sf], then the words [sfɪn] and [sfɪt] are not at all acceptable, despite [sfir] being a word of English with meaning. There are other such clusters in English, such as the [skl] in *sclera*.

English speakers will seem to have gradient reactions to these clusters, ranking certain clusters as 'more' English than others (Algeo, 1978). However, these reactions can still be attributed to categorical decisions from other rules. Algeo (1978) notes specifically that [sf] and [sv] belong to this category of existing but illicit clusters, but also that [sv] violates the additional rule of voicing agreement. If we consider voicing agreement a separate category from the one that [sf] and [sv] share, then while speakers will vary among their intuitions, and a single speaker will have consistent intuitions about each of these clusters, we can also consider that this voicing disagreement will make [sv] 'less' acceptable across the sum of speaker judgments. These phones can be held over from loan words or sounds that simply are no longer productive in English; [ʒ] and [ð] are two such singleton phones, rather than clusters, that are not legal in novel English words, but are still familiar to speaker due to their place in common words. Other entire sequences may have once occurred, such as [lg], but evolved out of the lexicon for one reason or another, such as vocalization (Algeo, 1978). In a sense, these phones and clusters can be considered 'exceptional'.

2.8 Future work

Here, I finished the matrix of feature specificity to frequency information seen in figure 2.1. In doing so, I uncovered some areas where further investigation can improve statistics for comparative purposes.

In particular, the BUFIA evaluation metric is not easily comparable with the statistics used for the cluster and bigram models, since the cluster and bigram models compare *presence* and frequency with acceptability, and BUFIA focuses on *absence*. It can be investigated whether the complement of the cluster and bigram lists provides more comparable results. Payne (2024) has developed a variant of BUFIA that outputs *positive* constraints, which could both improve the interpretability of statistics and find additional ways to improve BUFIA to account for more cases.

Secondly, it was noted that bigrams may cause an issue due to the flexibility of sonority in Polish. It would be beneficial to look more deeply into how sonority works in Polish to see if there are simple ways to improve this model to account for this without increasing the size of substrings, and thus the complexity of any learning problems.

Third, it would be interesting to expand this to other languages with similarly complex clusters, or other phonotactics questions. As well as Durvasula (2020) showing that a categorical approach to English clusters is as effective as a gradient one, BUFIA is being used to examine tone and other phonological features. Finding other ways to account for learning problems, whether phonological in nature or otherwise, in a categorical manner rather than a gradient one can greatly reduce the amount of data needed for learning problems and better inform how language acquisition occurs.

2.9 Conclusions

Here, examined six distinct phonotactic leanring models based on whether the representation were cluster-based, bigram-based, and feature-based, and whether they received categorical or gradient data as input. Broadly, I found that categorical data performs as well as, if not better than, gradient data in Polish. Additionally, I find that using categorical feature constraints better accounts for the clusters in Polish than the constraints derived by a gradient system, which finds a set of 100 constraints that fits the data worse than a smaller set of 50 constraints.

These results suggest a categorical approach to phonotactic constraints at the segment level. This is bolstered by the observation that HWPL is outperformed by both feature-based and segment-based bigrams (Albright, 2009). Since phonotactic learning has largely been gradient up to this point, these results suggest that constraints over attested clusters need a more categorical mechanism which may be distinct from the process that recognizes acceptable or unacceptable clusters. Further research would investigate the distinction between the two, which would better address how non-productive but attested clusters are stored in the lexicon.

3: MORPHOPHONOLOGY

3.1 Introduction

When building inflectional models, we are interested in identifying what information is needed to trigger particular changes. One goal of generative phonology is to categorize processes in a way that can be extended to novel data whenever possible, rather than listing off input and output forms. In morphophonology in particular, the difference can be seen in productive processes as opposed to exceptional forms. The English past tense /-d/ is a typical example, agreed to be the standard morpheme applied to novel words, though there are numerous exceptional processes or forms in isolation that must be memorized. For example, the copula 'to be' becomes 'was' in past tense, and 'to go' becomes 'went'. These are forms that must be memorized, with no single generalizable pattern. Other patterns, like the ablaut in 'sing-sang-sung', can be more difficult to classify and may be prone to variation because of a lack of regular pattern.

This problem can extend to items that once had a historical purpose that has since phased out - in this case, Polish yers (Gussman, 2007; Kraska-Szlenk, 2007). Historical yers (sometimes also called jers) come from a super-short vowel that, as seen in chapter 2, can result in a complex consonant cluster, as in [mgwa] ('fog', nom.) or [jabwkɔ] ('apple', nom.). However, they also can be elicited through morphological alternations, as in [v^{j}ɛɕ] ('country', nom.) and [fɕi] ('country', loc.). It is assumed that than the vowel in [v^{j}ɛɕ] underlies the lemma form, rather than being inserted to avoid the consonant cluster remaining when the suffix [i] is removed from [fɕi]. Kraska-Szlenk (2007) notes that the nominative is the least marked form, so when a yer is present, it always appears in the lemmatized nominative form.

Because of the yers' origins, their appearance in some token's underlying form is assumed to be memorized. However, yer alternations occur in a small fraction of word types from the Polish language, leading to an issue of data sparsity. This raises the question - when Polish speakers learn words with yers, are they in fact memorizing every form, or is there a generalizeable pattern? Moreover, is this a pattern that human speakers are able to learn that cannot be replicated by computers due to this sparsity, or potentially vice versa?

This issue is more relevant when considering words that do not have Polish origin, like [alabastɛr] ('alabaster') or [pɛrwa] ('pearl'). In the first case, inflecting from the nominative to the locative [alabastrax] shows that the final vowel in the nominative form is deleted, and thus, that a vowel in a foreign word can be treated like a yer. Even more striking, when inflecting [pɛrwa] to the genitive plural [pɛrɛw] or diminutizing to [pɛrɛwka], we see that a wholly new vowel is inserted and treated like a yer without being present in the input. This suggests that speakers have internalized some pattern, and calls into question how it might be learned.

In this chapter, I looked at a corpora of Polish words, with and without yers, to determine if there are particular environments that co-occur with yers, which speakers can use when deciding whether to label a vowel as a yer or not. Through a series of computational experiments, I use machine learning techniques to observe the environments and combinations of features used for yer prediction. In doing so, I create data sets that have the immediate environment of the yer written as single phonemes and as phonological features, as well as morphological information for each form, and investigate whether the features or phonemes allow for simpler generalizations.

Additionally, I compare decision tree algorithms to several different feed-forward neural networks, known as perceptrons. I organized these perceptrons by permuting their parameter settings to which settings will enable the perceptron to outperform the simpler decision tree. After introducing the data used, I begin by describing the results of the decision trees, then the perceptrons. I describe and compare the generalizations seen in both sets of models to investigate their commonalities, as well as where individual strengths aid each kind of model. Finally, I discuss what implications these have for similar machine learning problems.

3.2 Previous work

Much work on yers focus on how various representations can replicate the atypical insertion/deletion pattern. Historically, prior to deletion of yer vowels, Slavic languages like Polish favored open syllables (Kraska-Szlenk, 2007). The later preference toward deletion resulted in the clusters described in chapter 2. Kraska-Szlenk (2007) states that, historically, yers were "realized when another year (sic) followed in the next syllable, which was in turn deleted."

The modern generalization is rephrased by Rubach (2016) such that the yer "alternates with zero when a vocalic suffix is added to the stem", and Scheer (2018) adds, "yers appear on the surface if and only if the following vowel is also a yer." This results cyclic patterns of selecting which vowel surfaces (Lightner, 1965). Given the myriad available representational theories, the question became which representation best generalized the facts.

One theory in Polish (Szpyra, 1992) and Russian (Yearley, 1995) that yers exist to break up the complex clusters that resulted from their deletion. Using the Whole Morpheme Hypothesis, Gouskova and Becker (2013) attribute this alternation to a ban on mid vowels, though they note this does not apply to all mid vowels. This approach requires a lexical separation between morphemes that have a mid vowel that *can* be deleted (in essence, a yer), and those that have mid vowels that *cannot* be deleted. This approach has shown to be controversial, as it departs from from generalizations suggested above and instead requires heavy diacritization.

Instead, other theories prefer a representation that relies on the moraic status of yers. Scheer (2018) and Rubach (2016) both instead reference the empty nucleus of the yer, which is what makes it the target of the deletion rules stipulated above. They argue that this approach better generalizes between Slavic languages refute the idea that diacritization occurs on a morpheme rather than a segment (for example, a vowel with or without moras), instead marking a yer as having no mora and being assigned one at vocalization (Rubach, 2016). This raises the question of characteristic differences between surfacing yers and non-yer vowels, since mora assignment could potentially neutralize the contrast. Beňuš (2012) examines the phonetic status of yer vowels in Slovak to support the moraic view, determining some slight phonetic differences between the two categories but questioning whether these are perceptable to average listeners and, thus, learnable.

Approaches that rely on paradigmatic information, then, are beneficial in that there is additional contextual information that can bolster a pattern being learned. The Tolerance Principle (Yang, 2016) hinges on the idea that irregular morphological alternations can become productive when a sufficiently larger subset of data becomes a dominant pattern. Jarosz (2005) examines yers through the lens of Contextual Correspondence, which penalizes word forms that vary across a pardigm, in balance with Contextual Identity, which supports affix selection.

Kraska-Szlenk (2007) instead provides a statistical breakdown of the various word forms that exhibit yer alternations and what the morphophonological environment surrounding the yer looks like. Here, the argument is that a sufficiently robust pattern can be learned through analogy to develop into a regular alternation process, with Kraska-Szlenk noting that that children used the neologism ⟨komputra⟩ (with yer deleted, as opposed to ⟨komputera⟩) to support awareness of the paradigmatic information.

3.3 Language data

Polish data was used, with a specific focus on words with yer alternations. I used Unimorph (Kirov et al., 2018; Saloni et al., 2015; Woliński and Kieraś; Calzolari et al., 2016) to build paradigms based off words that shared a lemma form. Unimorph is a morphological annotation schema used for natural language processing tasks, which conveniently lays out each inflected word alongside its lemma form and morphological information. This schema made it simple to extract specifically the nouns and group them by lemma to track alternations across cases.

After grouping words by lemma to form a word paradigm and converting from orthography to IPA, a simple algorithm was used to annotate each word form as having a yer, based on key differences from the lemma form. This annotation separated the data into groups the machine learning algorithms would have to learn to distinguish. I compared every inflected form to the root in a character-by-character fashion. Words were marked as having a yer if, after removing a word-final vowel, the same position had a vowel in one root and a consonant in the inflected form.[1] For example, in the word pair [aniɔw→**ɛ**k] ('little angel', nom.) and [aniɔw→**k**ax] ('little angel', instr.), the right arrow and bolded character shows the point where the algorithm would decide there is a yer based on the *presence* of the vowel in the nominative and *absence* in the instrumental.

This followed from the observations by Kraska-Szlenk (2007) and Rubach (2016), where a yer will surface if the next vowel is not in a word-final open syllable, and will delete if the next

[1] This observation fails in the case of words with syncretized paradigms, such as ⟨człowiek⟩ ('boy', or [t͡ʃwɔvʲɛk] in IPA) or ⟨tydzień⟩ ('week', or [tɪd͡ʑɛɲ]), which inflect to ⟨ludzie⟩ ([lud͡ʑi]) in the plural and ⟨tygodnie⟩ ([tɪgɔdɲɛ]) in all forms but nominative respectively. Thus, words listed in Unimorph with either of these as the lemma were thus removed from consideration. Words that used these as a root, such as ⟨nadczłowiek⟩ ('Superman') were also removed.

vowel *is* in a word-final open syllable. This is best exemplified through chained diminutives: the suffix /-ɛk/ is used to mark diminutives, and the vowel in this suffix is a yer. Not only are speakers able to double diminutives (i.e., append /-ɛk/ twice to a root), a feminine noun is also able to additionally take the female suffix /-a/. To illustrate, consider the following surface forms of the various diminutizations of the word 'kot', meaning 'cat':

1. kɔt + a → kɔta[2]
2. kɔt + ɛk → kɔtɛk
3. kɔt + ɛk + a → kɔtka
4. kɔt + ɛk + ɛk → kɔtɛt͡ʃɛk[3]
5. kɔt + ɛk + ɛk + a → kɔtɛt͡ʃka

Data were annotated in two ways. In the first, words were annotated for having either a *form* yer if the specific word being considered had a yer, or a *paradigm* yer if some word in the paradigm did, but not necessarily the current form. For example, [vʲɛɕ] would be annotated as having both, but [fɕi] would only have the paradigm yer. Secondly, the phonological information on either side of the potential yer was presented as a *symbol*, a *feature list*, or a *combination* of the two. If there was no yer in a given word, this information would come from the final closed syllable in the word, saving the phone immediately proceding and following the vowel. This was used to test the learning algorithms' ability to generalize features to find environments where yers tended to appear.

Additionally, a copy of this data set was created where diminutive forms were removed. A word was considered to be a diminutive if 1) it has [ɛk] or [ɛt͡ʃɛk] as a suffix of the string, and 2) somewhere in Unimorph there is a form that is identical to the current form *without* either [ɛk] or [ɛt͡ʃɛk]. This targets the lemma of each paradigm, though diminutized items are listed in Unimorph under the diminutive form - for example, inflected forms of /kɔtɛk/ are not listed under /kɔt/, even though the diminutive is itself inflected. Note that the presence of the diminutive suffix can induce a secondary phonological change in the base word, namely palatalization of the final consonant, that must be taken into account to align the undiminutized and the diminutized forms.

[2] Here, the /-a/ suffix signifies the genitive form of the masculine noun, rather than the feminine. However, this form demonstrates that there is no interaction between the final /a/ and the preceding vowel /ɔ/, so I include it here.

[3] Here, the [t͡ʃ] cluster results from the palatalization of the [k], which is irrelevant to the current problem.

	Dim.	No Dim.
Total	91893	89529
Global	4127	1765
Local	639	233

Table 3.1: Number of tokens in each data set, as well as counts of relevant yer parameters.

3.3.1 Corpus statistics

There were 91893 total word forms in the corpus I extracted from Unimorph, with each item marked for 1) if a yer surfaced locally in this word, labeled Form_Yer, 2) if a yer surfaced globally somewhere in this item's inflectional paradigm, labeled Paradigm_Yer and 3) relevant morphophonological information as previously described. The table in 3.1 shows how many items were in each of the two data sets (one with diminutives, one without), how many of those items had paradigmatic 'global' yers, and how many had surfacing 'local' yers; items with local yers are inherently a subset of items with global yers, since the global yers are determined by items with surface yers somewhere in the paradigm.

3.3.2 Experimental design

These experiments are classification tasks to determine whether, given the environment in a word where a yer alternation would occur, a yer would surface based on the immediate phonological environment and morphological information about the word. Different models were trained according to the following combinations of the factors:

- Whether the model predicts the Form_Yer or Paradigm_Yer, as well as a Form_with_Paradigm case
- Whether single symbol consonants or consonant features are used, or both
- Whether or not diminutives are included

Testing how machine learning techniques predict a Form Yer compared to a Paradigm Yer would investigate whether the information provided was sufficient to predict either of these patterns, and the differences in them. The Form with Paradigm case is added to see how paradigmatic information can inform pattern-learning in addition to environmental information[4]

[4] The opposite, Paradigm with Form, naturally is redundant because knowing a yer is in a given form necessitates it being in the larger paradigm.

- if a yer alternation is known to occur in a given paradigm, will this particular word surface with a yer? The FWP models are not provided information for what form (nominative, genitive, etc.) the yer surfaces, only that the alternation occurs somewhere. By giving the model access to this information, I explore whether and how these learning paradigms leverage it. For example, does this information improve a decision made by these models, or is the morphophonological information in a single word sufficient for the models to accurately predict yer placement?

Because of how few items surface with yers, as seen in table 3.1, the training set and test set were identical, so that the models could consider as many tokens as possible. Only the immediate phonological environment was listed as variable, which leads to significant competition from non-yer items that already massively outnumber items with yer. Since the research question is interested in the environments chosen by these models to determine ability to generalize, overfitting to data is not a concern.

The input for each word form included global or local yer information (or both, depending on the model being generated), as well as the symbol for the phone to the immediate left and right of the potential yer position. The features for each of these symbol were also listed, according to Hayes (2011), as either binary +/- or ∅ where appropriate. The output classified each item as either having a global or local yer, per the model specifications, or not having a yer.

3.4 Decision trees

Decision tree models were used to observe what combination of features were most relevant to determining an underlying yer. In a decision tree, data is subdivided into sets depending on categorical questions about features in particular orders (Flach, 2012; Duda et al., 2001). Depending on the implementation of the decision tree algorithm, there is some function that weighs features to determine which ones play the strongest role in decision making, depending on an internal scoring mechanism. Once a feature is determined to be sufficiently predictive, that subset of data is further subdivided based on additional information, with the end goal of selecting groups of features that will accurately classify and label input items. These results are then easily interpretable by users to make generalizations about results (compared, for example, to neural networks, which are infamous for being 'black boxes' and difficult to interpret (Balkir et al.,

	Precision	Recall	F-Score
Paradigm/Combined/Dim	0.901	0.548	0.682
Paradigm/Feature/Dim	0.904	0.558	0.690
Paradigm/Symbol/Dim	0.886	0.543	0.674
Form/Combined/Dim	0.886	0.097	0.175
Form/Feature/Dim	0.909	0.125	0.220
Form/Symbol/Dim	0.841	0.083	0.151
Paradigm/Combined/No dim.	0.983	0.529	0.681
Paradigm/Feature/No dim.	0.983	0.529	0.686
Paradigm/Symbol/No dim.	1.00	0.467	0.637
Form/Combined/No dim.	0.907	0.292	0.401
Form/Feature/No dim.	0.925	0.369	0.486
Form/Symbol/No dim.	1.00	0.313	0.462

Table 3.2: Precision and recall scores, only for items that are marked as having either a local or a global yer.

2022; Yeh et al., 2022; Danilevsky et al., 2020; Doshi-Velez and Kim, 2017)). Decision trees have been used linguistically for determining phonetic output based on sociolinguistic factors (Eddington, 2010) or for determining contrast in a language's phonemic inventory (Chandlee, 2023).

However, the structures of decision trees are also somewhat unstable, and the introduction of new data can trigger large reorganizations (Duda et al., 2001). If parameters have a considerable number of features (for example, the symbol features I used), this can also cause a very flat tree structure. Conversely, trees with many splits must be translated to complex rules. While still interpretable, this can make predictions difficult to generalize if each rule maps to a singular form.

I used WEKA (Frank et al., 2016), a machine learning software managed by the University of Waikoto that compiles many learning algorithms for efficient comparison. Specifically, I used the J48 decision tree algorithm, which is WEKA's open source Java implementation of the C.45 classification tree (Quinlan, 1993). This is a 'classification and regression tree' (or CART) style algorithm (Breiman et al., 1984), which describes the cost function used to decide which features become decision points.

3.4.1 Results

In table 3.2, models are listed alongside their precision and recall scores to reflect their effectiveness. The F-scores, or harmonic means of the precision and recall, are also listed for easy comparisons. Precision scores, ensuring that the items selected by the model are in fact yer-bearing, remain high. The recall scores, grading how many of the yer-bearing items are selected, drop significantly. Decision trees are more effective at accurately predicting whether a word has a paradigm yer than a form yer. The paradigm models were only able to recall about half of the yer-bearing items, which means that the model still struggles in finding a robust pattern that easily describes an environment where yers will occur.

Form models perform worse than the paradigm model. This makes sense because these models have a more complicated task. They are asked to determine both 1) is there an underlying yer in this word, and 2) does this word satisfy the environment that would make this yer surface. Precision scores are still high - the items selected generally do have yers - but because the recall scores are so low, very few of the yer-bearing items are being selected. The confusion matrices, seen in appendix **??**, are grids that have along one axis what an item was predicted as and along the other, its correct label (here, whether or not it had a yer), with each position being the number of items sorted in that grouping. These matrices clarify that the ratio of high precision to low recall is because the form models are selecting very few items as having yers generally, shown by low numbers in the 'labeled True' column. As shown in table 3.1, 639 items had a local yer; for all form models, fewer than 100 items were selected to be yer bearing. This is partly a consequence of the fact that words with yers are such a minuscule part of the Polish lexicon, and by ignoring all items that do have yers, the accurately-labeled words *without* yers will increase the weighted scores. In this, these models fail to make the generalizations the paradigm model make, as well as failing to find the patterns for surfacing.

3.4.2 Models with and without diminutives

After excluding diminutives from decision trees (results in table 3.2), we observe that the paradigmatic models have a very slight decrease in overall F-scores. The precision increases, meaning that it improves at selecting only target items, with a minor decrease in recall. More sizeable changes are seen for the form models, where F-scores more than double for all three models. The precision increase is minor for all but the Form/Symbol/No Diminutive model, which *only* selects target items, but the recall score increases significantly for all three. Note that the recall score is *still low*, selecting less than half of the target items despite this increase. The confusion matrices again confirm that the low recall is an artifact of the ratio of data; despite removing over 2000 items that had a yer, most of these had been incorrectly predicted by the diminutive-inclusive model as not having a yer. In fact, the form models using symbols is the only one of the three to accurately predict *more* items with yers instead of *fewer*.

Since including diminutives made the form models perform worse, it can be surmised that these models were unable to generalize the diminutive suffix (marked in data as having a [k] immediately following the yer position). In the dataset including yers, there were 8154 words with a right-environment [k] (regardless of diminutive status), of which 2362 had paradigm yers and 406 had form yers. After removing diminutives, there are no right-[k] forms that have yers, but 5792 non-yer words remain. Because this is higher than the total number of yer-carrying words in the diminutive inclusive list, the diminutive pattern is simply too infrequent and is overwhelmed by to combat the non-diminutive [k] without additional information not provided here.

The paradigm models, on the other hand, did not change much, based on the precision and recall values and near-identical F-scores. However, closer examination of the tree structure found that the models that included diminutives (seen in **??** and **??**) are much more intricate compared to ones that did not (**??**). Though the scores were similar, the difference in tree structure shows that the decision tree algorithm found considerably more information when working to exclude items with [k] in the right environment that were not diminutives.

3.4.3 Using paradigmatic information as a parameter of form-based models

In table 3.1, we can see broadly that, while precision drops in some cases, there is a significant increase in recall scores, which results in higher F-scores for all models. This means that, compared to non-FWP models, introducing paradigmatic information means that the models are able to accurately select most of the target items as having yers, with a much smaller increase in

	Precision	Recall	F-Score
FWP/Combined/Dim	.891	.958	0.923
FWP/Feature/Dim	.891	.958	0.923
FWP/Symbol/Dim	.862	.984	0.919
FWP/Combined/No dim.	.933	.893	0.861
FWP/Feature/No dim.	.930	.906	0.870
FWP/Symbol/No dim.	.826	.927	0.887

Figure 3.1: Precision and recall scores for the FWP models.

the number of items incorrectly selected.

This addresses the issue of target items for the form model having inadequate information. If a word has a paradigm yer, this implicitly provides that the word at hand has the correct syllable structure, appropriate cluster sequences, and so on. The model need only determine if the information provided can determine whether a suffix follows that would block the yer from surfacing.

We also see that removing items with diminutives here *decreases* recall and F-scores, converse to the plain form models. While the presence of diminutives only served to confuse the form model, with inadequate information to differentiate a diminutive right-[k] to a non-diminutive one, access to the paradigmatic information is sufficient to clarify this: if there is a right-[k] and the paradigm has a yer, it *must* be part of a diminutive structure.

3.5 Perceptron model experiments

With neural network models making large strides in the field of natural language processing, it makes natural sense to compare the decision trees against a series of neural networks, to compare and contrast their strengths on a data set like this - namely, one where the target items form a small subset of the data. Using the MultiLayerPerceptron package in WEKA (Frank et al., 2016), I built additional models for yer prediction using perceptrons (Gallant, 1990), which are an early ancestor of the modern neural network. This means that it has a simple architecture, though it does include backpropagation.

With the perceptron models, I manipulated two key parameters - training time and learning rate. The training time referred to how many ‘epochs’ the perceptron ran for, ostensibly giving it more exposure to the training data. The learning rate, on the other hand, refers to how much the weights may change during each learning iteration (Biehl and Schwarze, 1995). A high learning rate may make a model more sensitive to small changes, but also can make it unstable. I used the same six model configurations, each with and without diminutives, as seen for decision trees in table 3.2. Each model was tested at three training times (after 5 epochs, 10 epochs, and 50 epochs) and three learning rates (.10, .30, and .50). This configuration resulted in 108 models, of which 10 failed to learn any pattern and are marked with a question mark '?' in tables 3.3 and 3.4. An additional 54 perceptrons were run for the FWP models, resulting in a total of 162 perceptrons to examine.

3.5.1 Neural network details

The MultilayerPerceptron package was the basic foundation for all neural network models. The only parameters manipulated were learning rate and training time. The number of hidden layers was left to the default setting, which averaged the number of attributes and classes. Attributes refer to how many variables were considered per model - the symbol models had 5 attributes (case, plurality, and the phone preceding and following the yer position, as well as whether that item had the target yer), and the feature model had 47 (case, plurality, yer presence, and all features for both the left and right immediate consonant). Naturally, the 'combined' models had all attributes from both attribute sets, and the FWP models had both the paradigm yer and form yer as attributes. Classes were the possible values for all attributes - the yer attributes were binary, feature attributes were binary except when a value of ∅ was appropriate (in which case, three classes), plurality had two (singular and plural), and morphological case had 7. When form models were used, the immediately preceding environment had 39 options, and the immediately following had 26.

3.5.2 Results

Precision for all diminutive models averaged at 88.7%, and for non-diminutive models averaged 96.6%, meaning that the items the models selected as yer-bearing generally did have yers. However, the recall was significantly lower, with paradigm models (both diminutive and non-diminutive) averaging 52.9%, diminutive forms models 10.2%, and nondiminutive form models

	5 Epochs		10 Epochs		50 Epochs	
Learning Rate: .10	Prec	Rec	Prec	Rec	Prec	Rec
Paradigm/Combined/Dim	0.999	0.528	**0.811**	**0.661**	0.854	0.630
Paradigm/Feature/Dim	**0.998**	**0.544**	0.808	0.596	0.806	0.656
Paradigm/Symbol/Dim	0.996	0.530	0.850	0.622	**0.876**	**0.620**
Form/Combined/Dim	?	0.000	1.000	0.045	1.000	0.114
Form/Feature/Dim	?	0.000	1.000	0.055	0.858	0.180
Form/Symbol/Dim	?	0.000	0.966	0.044	0.848	0.166
Paradigm/Combined/No dim.	0.999	0.528	**0.997**	**0.556**	**1.000**	**0.567**
Paradigm/Feature/No dim.	**0.998**	**0.544**	0.998	0.550	0.998	0.559
Paradigm/Symbol/No dim.	0.996	0.530	0.996	0.534	0.996	0.558
Form/Combined/No dim.	?	0.000	0.941	0.275	1.000	0.114
Form/Feature/No dim.	?	0.000	0.929	0.279	0.944	0.361
Form/Symbol/No dim.	?	0.000	?	0.000	0.986	0.300
Learning Rate: .30	Prec	Rec	Prec	Rec	Prec	Rec
Paradigm/Combined/Dim	0.777	0.602	0.781	0.548	0.823	0.585
Paradigm/Feature/Dim	**0.998**	**0.531**	0.813	0.591	0.761	0.632
Paradigm/Symbol/Dim	0.996	0.529	**0.780**	**0.663**	**0.842**	**0.616**
Form/Combined/Dim	0.874	0.326	1.00	0.036	1.00	0.105
Form/Feature/Dim	0.893	0.322	1.00	0.075	0.859	0.182
Form/Symbol/Dim	?	0.000	1.00	0.028	1.00	0.092
Paradigm/Combined/No dim.	1.000	0.486	1.00	0.528	**0.985**	**0.578**
Paradigm/Feature/No dim.	1.000	0.486	0.954	0.537	0.996	0.568
Paradigm/Symbol/No dim.	**0.996**	**0.529**	**0.997**	**0.535**	1.00	0.547
Form/Combined/No dim.	0.874	0.326	1.00	0.300	0.944	0.361
Form/Feature/No dim.	0.893	0.322	1.00	0.296	0.943	0.356
Form/Symbol/No dim.	?	0.000	1.00	0.236	1.00	0.296
Learning Rate: .50	Prec	Rec	Prec	Rec	Prec	Rec
Paradigm/Combined/Dim	0.696	0.614	0.869	0.453	1.000	0.197
Paradigm/Feature/Dim	0.670	0.660	0.872	0.421	0.938	0.377
Paradigm/Symbol/Dim	**0.771**	**0.658**	**0.872**	**0.460**	**0.721**	**0.682**
Form/Combined/Dim	1.000	0.002	1.000	0.011	0.973	0.113
Form/Feature/Dim	1.000	0.003	1.000	0.056	0.943	0.130
Form/Symbol/Dim	?	0.000	1.000	0.003	1.000	0.086
Paradigm/Combined/No dim.	0.971	0.527	0.990	0.511	0.998	0.519
Paradigm/Feature/No dim.	**0.934**	**0.562**	**0.977**	**0.561**	**0.916**	**0.601**
Paradigm/Symbol/No dim.	0.996	0.530	0.996	0.534	0.992	0.539
Form/Combined/No dim.	0.890	0.348	0.920	0.343	0.944	0.361
Form/Feature/No dim.	0.971	0.292	0.940	0.339	0.944	0.361
Form/Symbol/No dim.	1.000	0.236	1.000	0.240	1.000	0.300

Table 3.3: Precision and recall for perceptron models trained with learning rate of 0.10. Cells marked with '?' are undefined due to recall of 0. Bolded values are models with the highest harmonic mean of precision and recall, per figures in 3.4.

	5 Epochs			10 Epochs			50 Epochs		
Dim.	.10	.30	.50	.10	.30	.50	.10	.30	.50
P/C	0.691	0.678	0.653	**0.728**	0.644	0.595	0.725	0.684	0.330
P/F	0.704	0.693	0.665	0.686	0.685	0.568	0.724	0.690	0.537
P/S	0.692	0.691	**0.710**	0.718	0.717	0.602	**0.726**	0.711	0.701
F/C	?	0.003	0.003	0.087	0.069	0.022	0.205	0.190	0.202
F/F	?	0.006	0.006	0.104	0.140	0.107	0.298	0.300	0.228
F/S	?	?	?	0.084	0.055	0.006	0.277	0.169	0.159
No Dim.									
P/C	0.691	0.654	0.684	0.714	0.691	0.674	0.724	**0.729**	0.683
P/F	**0.704**	0.654	0.702	0.709	0.687	0.713	0.716	0.723	0.726
P/S	0.692	0.691	0.692	0.695	**0.764**	0.695	0.715	0.707	0.698
F/C	?	0.475	0.500	0.425	0.462	0.500	0.522	0.522	0.482
F/F	?	0.473	0.449	0.429	0.457	0.498	0.522	0.517	0.522
F/S	?	?	0.382	?	0.382	0.388	0.461	0.457	0.462

Table 3.4: F-measures for all Form and Paradigm perceptron models. Cells marked with "?" are undefined due to recall of 0. Bolded values are the highest F-score per

from 32.5%, suggesting that these models acquired at most about half of the target yer items.[5]

Broadly, these generalizations hold to the perceptron models, but there are a number of additional specifics to address. Firstly, at lower training times and learning rates, the models are unable to learn *any* yer patterns - in particular, the form models at 5 epochs (learning rate .10) as well as the Form/Symbol/No-diminutive model after 10 epochs (learning rate .10), both Form/Symbol models after 5 epochs (learning rate .30), and the Form/Symbol/Diminutive model after 5 epochs (learning rate .50).

Secondly, it holds that precision is consistently higher than recall values. The only three models where recall is equal to or greater than perception are FWP/Feature (with diminutives, training time of 50 epochs, and learning rate of 0.30) and two FWP/Combined models (with and without diminutives, training time of 10 epochs, and 0.50 learning rate), which are marked in tables 3.1 with italics. This suggests that these perceptrons tend to limit their generalization to minimize the number of items without yers they select. Form models in particular have a large difference between precision and recall. Note that these values are weighted specifically to the items with yers, and that models that fail to select *any* items with yers nonetheless have a

[5] It is difficult to suggest that these models performed 'at chance', since these yer-bearing items are a small subset of the data. As shown in table 3.1, the corpus with diminutives has about 91000 words, of which about 3000 have paradigm yers and 600 have form yers. The corpus without diminutives has about 89000 words, of which about 1500 have paradigm yers and 300 have form yers.

weighted average accuracy of .90 and higher. In comparison, table 3.1 shows that, when using decision trees, all three diminutive FWP models have higher recall than precision, as well as the symbol model without diminutives.

3.5.3 Models with and without diminutives

The effect of removing diminutives is similar with perceptrons as compared to decision tree models. In all three form models, we see an increase in recall for form models when diminutives are removed, though not to a level that is competitive with paradigm models.

Broadly, these models are competitive with the decision tree models, though at particular parameter settings. With diminutives included, there was significant variety in how the perceptrons performed - with F-scores anywhere from 0.003 to 0.3 (of the models that selected any target items), as seen in 3.4. The range for models without diminutives is from 0.382 to 0.522 which, which does suggest that the diminutives are similarly causing instability in the model's ability to generalize.

3.5.4 Using paradigmatic information as a parameter of form-based models

In tables 3.6, we see performance of perceptrons when given paradigmatic information to predict form yers, which parallels the findings in 3.1. As with the decision trees, we see that recall in particular rises for all models, resulting in overall higher F-scores.

However, we additionally see an issue with the FWP/Combined/Diminutive models excessively rejecting target items after 10 training epochs. For decision trees, the corresponding model tied with the FWP/Feature/Diminutive for best model performance, with the caveat that the Combined decision tree excluded non-feature information, thus exactly matching the Feature tree. In terms of perceptrons, the FWP/Combined/Diminutive model outperforms for the 5-epoch, .50 training rate and 10-epoch, .10 training rate models, but is outclassed by FWP/Feature/Diminutive elsewhere.

	5 Epochs		10 Epochs		50 Epochs	
Learning Rate: .10	Prec	Rec	Prec	Rec	Prec	Rec
Combined/Dim	?	0.000	**0.919**	**0.876**	0.976	0.637
Feature/Dim	?	0.000	0.979	0.673	0.951	0.818
Symbol/Dim	?	0.000	0.951	0.767	**0.942**	**0.784**
Combined/No dim.	?	0.000	0.919	0.876	**0.938**	**0.910**
Feature/No dim.	?	0.000	**0.916**	**0.893**	**0.938**	**0.910**
Symbol/No dim.	?	0.000	0.903	0.876	0.926	0.918
Learning Rate: .30	Prec	Rec	Prec	Rec	Prec	Rec
Combined/Dim	0.933	0.717	**0.959**	**0.776**	0.993	0.455
Feature/Dim	0.953	0.737	0.979	0.660	***0.923***	***0.923***
Symbol/Dim	**0.959**	**0.807**	0.992	0.571	0.985	0.624
Combined/No dim.	0.933	0.717	**0.944**	**0.863**	0.945	0.888
Feature/No dim.	0.911	0.747	0.979	0.617	**0.942**	**0.910**
Symbol/No dim.	**0.959**	**0.807**	0.997	0.619	0.922	0.918
Learning Rate: .50	Prec	Rec	Prec	Rec	Prec	Rec
Combined/Dim	**0.923**	**0.876**	***0.897***	***0.926***	0.955	0.066
Feature/Dim	0.976	0.574	0.960	0.781	**0.952**	**0.812**
Symbol/Dim	0.947	0.757	0.927	0.859	0.976	0.710
Combined/No dim.	**0.923**	**0.876**	***0.911***	***0.923***	0.912	0.884
Feature/No dim.	0.917	0.811	0.933	0.897	0.926	0.918
Symbol/No dim.	0.954	0.803	0.933	0.897	**0.934**	**0.914**

Table 3.5: Precision and recall scores for the FWP models, with a learning rate of 0.10. Cells marked with '?' are undefined due to recall of 0. Bolded values are models with the highest harmonic mean of precision and recall, per figures in 3.6.

	5 Epochs			10 Epochs			50 Epochs		
Diminutives	.10	.30	.50	.10	.30	.50	.10	.30	.50
FWP/Combined	?	0.811	**0.899**	**0.897**	0.858	0.911	0.771	0.624	0.123
FWP/Feature	?	0.831	0.723	0.798	0.789	0.861	0.880	**0.923**	0.877
FWP/Symbol	?	0.876	0.842	0.849	0.725	0.892	0.856	0.764	0.822
No Diminutives									
FWP/Combined	?	0.811	**0.899**	0.897	0.901	0.917	0.924	0.916	0.898
FWP/Feature	?	0.821	0.861	**0.904**	0.757	0.915	0.924	**0.926**	0.922
FWP/Symbol	?	0.876	0.872	0.889	0.764	0.915	0.922	0.920	0.924

Table 3.6: F-measures for all Form With Paradigm (FWP) perceptron models. Cells marked with "?" are undefined due to recall of 0. Bolded values are the highest model in its group.

3.6 Analyses

3.6.1 Generalizations predicted by decision tree paths

In including the consonants that immediately precede and follow the yer position, as well as all of its features, I was able to look at whether phonological features or particular segments were sufficient for models to generalize well. As seen in 3.2, the differences between symbol and feature-based models in terms of recall and precision are miniscule.

Symbol models selected fewer environments and generally rejected entire possible contexts because it could not effectively generalize. However, the affricates /t͡ʃ/ and /t͡ʂ/ in the right environment were selected as yer-bearing contexts. In paradigmatic models, the right context /r/ and /k/ also were predictors, but there was some granularity here. The right [r] only predicting a yer in the paradigm in the case of the instrumental, essive, and dative plurals. The right /k/, however, was more complex; in §3.6.4 I elaborate on how these results relate to the diminutive. When the right /k/ is selected, the paradigmatic model will then examine the left context and, based on its contents, may use morphological information to predict a yer. The generalizations with final [k] are difficult - depending on which preceding consonant is chosen, the combination of true and false cases varies, though none of these select the accusative (barring [ʂ], the only preceding symbol that is marked yer-bearing with no additional criteria). If the tree branches further to look at plurality, then in general, the plural will be yer-bearing and the singular will not. The two exceptions to this are [w] and [n], which select the vocative and genitive singular (as well as instrumental, essive, dative again). This is the opposite of what is expected: in words that have the diminutive, it generally surfaces in the nominative (Kraska-Szlenk, 2007).

Feature-based (and combined) models add clarity by generalizing across contexts more effectively. Paths of particular interest, each predicting presence of either the paradigm or form yer, are highlighted here:

- Paradigm/{Feature, Combined}/No Diminutives
  - Right_Dorsal: +; Right_Sonorant: +; Left_Back: –; Left_Voice: –
  - Right_Dorsal: –; Right_Approximant: +; Case: INS; Plural: PL
  - Right_Dorsal: –; Right_Approximant: +; Case: ESS; Plural: PL

  - Right_Dorsal: –; Right_Approximant: +; Case: DAT; Plural: PL
  - Right_Dorsal: –; Right_Approximant: –; Right_Anterior: –; Right_Continuant: -
  - Right_Dorsal: ∅

- Form/{Feature, Combined}/No Diminutives and Form/Combined/Diminutives
  - Right_Dorsal: ∅; Plural: SG; Case: ACC; Left_Anterior: –
  - Right_Dorsal: ∅; Plural: SG; Case: NOM

- Form/Features/Diminutives
  - Right_Dorsal: –; Right_Anterior: –; Right_Continuant: –; Left_Coronal: –
  - Right_Dorsal: –; Right_Anterior: –; Right_Continuant: –; Left_Coronal: ∅[6]
  - Right_Dorsal: ∅; Plural: SG; Case: ACC; Left_Anterior: +
  - Right_Dorsal: ∅; Plural: SG; Case: NOM

By and large, the most relevant morphological cases were nominative and accusative, which follows expectations. Per Kraska-Szlenk (2007), when a yer surfaces, it is almost always found in the nominative. The nominative is both the least marked form, and the nominative singular is least likely to include the open syllable suffixes that would prevent a yer surfacing (Kraska-Szlenk, 2007; Rubach, 2016). The accusative, often identical to the nominative, comes with the added requirement that the consonant to the left be negative for anterior, which includes most of the palatal obstruents and and affricates. In terms of plurality, the nominative and accusative always selected the singular. This often co-occured with right consonants marked 0 for dorsal which singled out the affricates: /t͡ʃ, t͡s, d͡ʒ/. Similarly, when the right consonant was non-dorsal but also not ∅, it was also marked for non-anterior and non-continuant, so long as the left consonant was non-coronal. This again selects the affricates (now also including /t͡ʂ, d͡ʐ/), as well as palatal stops and the palatal nasal.

[6] The only phone marked ∅ for coronal was a placeholder boundary symbol # for empty clusters, which only occurred in the non-yer words. WEKA also marked this as occurring in zero instances. As such, this case is disregarded.

3.6.2 Weighted nodes used for predictions in perceptron models

Looking at the perceptron outputs, we can compare the weight of given nodes between two models, though the result is less interpretable than the chains of features that can be inferred from decision trees. Nodes that are highly weighted contribute heavily to a perceptron's prediction. This would be equivalent to the features that a decision tree's cost function determines to be strong predictors, since the function determined that they were more important to the decision than other features. For example, when we compare the FWP/Combined/Diminutive models with learning rate of .50 at both 10 and 50 epochs, we're able to make generalizations about the node weights. These two models are good for quick comparison because one performs very well (F-score of 0.911), and one performs very badly (F-score of 0.123), with only a difference in training time, seen in table 3.6. Focusing specifically on the symbols immediately before the yer position, we can, for example, generalize that the average weight decreases from -0.1053 at 10 epochs to -0.5688 at 50 epochs. This means that, at 10 epochs the perceptrons attribute low effect from symbol data, most of which the model weakly negatively correlates with having yers. The only symbols that have positive weights are /j/ and /v^j/, at 0.2367 and 0.2025 respectively. After 50 epochs, all weights for preceding symbols have become more strongly negative; /j/ is still weakly positive at 0.0439, but /v^j/ is now moderately negative at -0.4098.

Most values in both models range between 1 and -1, and value beyond that range indicate higher correlations with those attributes. In the 10-epoch model, the only 'strongly' weighted attributes are 'accusative' (-1.5467), 'nominative' (-2.4095), 'plural' (-2.4095), and there being no paradigm yer (3.3267). The 50-epoch model, on the other hand, has more than triple this number, including a number of featural attributes:

- accusative (-5.1292)
- nominative (-4.8348)
- essive (2.0183)
- dative (1.9972)
- vocative (1.3596)
- genitive (1.2257)
- plural (5.5689)
- preceding /l/ (-1.1611)
- preceding /v/ (-1.0026)
- preceding non-dorsal(-1.0236)

- preceding lateral (-1.0881)
- preceding non-lateral (1.0079)
- following non-anterior (-1.3068)
- following delayed release (-1.1961)
- no paradigm yer (8.1073)

3.6.3 Generalizing with form

As previously mentioned, models predicting yers in a single word form (without paradigmatic information) perform worse than those predicting a yer somewhere in the paradigm. In all cases across decision trees and perceptrons, model recall is lower than the corresponding paradigm models, and especially lower than the FWP models. This highlights the need for paradigmatic information when making predictions about a word form.

However, with varying training times and learning rates, we see additional difficulties in the form models for perceptron. Recall, first, that there are only 233 tokens with a form yer in the diminutive data set, compared to over 1500 with a paradigm yer - from the beginning, the learning problem is more difficult, since the models must make predictions with much less data. At all learning rates, we see form-based models that fail to select any items, in turn failing to learn any generalization for yers. At the lowest learning rate of .10, no form model learns after 5 epochs, and even after 10 epochs the Form/Symbol/No-diminutive model still cannot learn. At a learning rate of .30, the Form/Symbol models still fail to learn after 5 epochs, though at 10 epoch all models have selected at least one item. With a learning rate of .50, the only model that fails is the Form/Symbol/Diminutive model at 5 epochs.

At low learning rates, this inability to learn from form alone complicates the FWP models, which otherwise tend to be the highest performing models. None of the FWP models learn after only 5 epochs at a learning rate of .10, though as soon as either the learning rate or training time increases, they're able to learn effectively. Since all of the paradigm models learned something, even with low recall, the form yer may be too small of a class to generalize at that time for perceptrons, even if all decision trees are able to learn *something*. Thus, perceptrons appear to be less robust than decision trees when faced with extremely sparse data.

In particular, all three FWP/Combined/Diminutive models show a sharp drop between training time of 10 epochs and 50 epochs in table 3.6; the Paradigm/Combined/Diminutive model

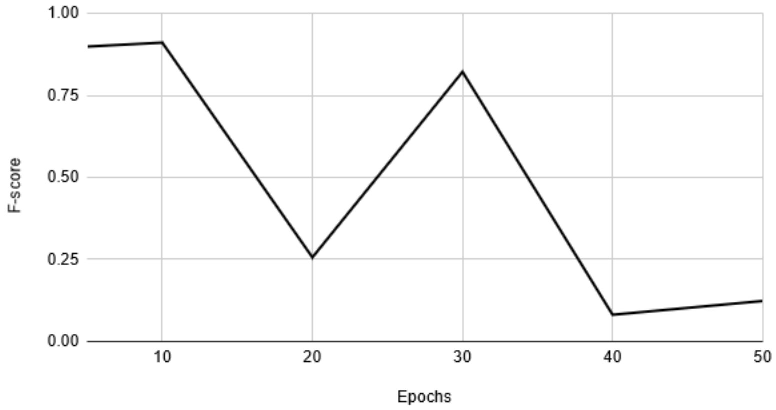

Figure 3.2: Comparing the F-measures of FWP/Combined/Diminutive model.

(learning rate 0.50) shows a similar drop. For example, in the FWP model with a learning rate of .50, it correctly selects 42 and fails to select 597. Compare this to the same model at .30, which selects 592 and fails to select 47. This shows a tendency to overly prefer rejecting yers at high training times, when these models had access to more data. To further investigate this, I ran three additional models at the 0.50 learning rate setting, using training times of 20, 30, and 40 epochs, with the F-measures listed in figure 3.2 for direct comparison. The FWP/Combined/Diminutive models at 0.50 goes from one of the best performing models (scoring higher that 93% of the models generated, including those in 3.4) to one of the worst (the lowest-performing FWP model that learned any pattern, and scoring lower than 83% of models, of which 10 learned some pattern and 16 did not).

When observing these models at intermediate learning stages - again, after 20, 30, and 40 epochs - we see that success in learning varies wildly. Figure 3.2 shows how the F-score of the FWP/Combined/Diminutive model (learning rate .50) oscillates wildly during training.

3.6.4 Effect of diminutives on learnability

Both decision tree models and perceptron models exhibit similar patterns in how they handle, or fail to handle, diminutive words. The difference in how the paradigm models compared to form models, for both decision trees and perceptrons, approach diminutives shows two things: how this dataset succeeds and fails to include information that would make diminutives detectable, and how decision tree structures are able to make themselves more complex to accommodate data that complicates a hypothesis in a way that is less apparent than perceptrons. For the first, the form and paradigm models ostensibly had access to the same information but were trained on different subsets of the data. Words targeted by the form model made up a *much* smaller part of the dataset, which likely exacerbated this issue. With the perceptrons in particular, we can notes how the paradigm models not only had higher F-scores in table 3.4, but also how, even with diminutives included, the range is generally smaller[7] (range of 0.189) compared to the form models with diminutives (range of 0.297). However, even when information was removed for the no-diminutive models, form models for both decision trees and perceptrons still lacked information needed to determine if a single form had a yer. If we follow the theory proposed by Rubach (2016), we will likely need more specific syllabic information for better performance. The paradigm models on the other hand had more data to draw from, but that data also seemed more likely to represent the necessary information.

Conversely, it could be that the question of 'does this word belong to a paradigm that has a yer' is more forgiving. Paradigm models had high precision, meaning few non-target items were selected, but the low recall meant they still overlooked many items. Remember that the diminutives were massively overlooked in the form models for decision trees due to the ratio of non-yer words with [k] in the right environment to words that had [k] in the right environment but *did* have yers. If the paradigm model found fewer situations with heavily unequal ratios, the decision tree algorithm may have been able to resolve the question by creating more classes. Compare the paradigm decision tree model in **??**, with a maximum depth of 16 levels, to that in **??**, with a maximum depth of 5. A closer statistical analysis of the results could further elucidate whether the data is more suited to paradigmatic information, whether the decision tree algorithm reacted more favorably with the ratio of items in the paradigm data set, or some combination of the two.

Of the three decision tree models in each of the form diminutive and form non-diminutive group, the best performing for the diminutive models had an F-score of 0.220 (Form/Feature, in table 3.2, and for non-diminutives had a score of 0.486 (again Form/Feature, in the same table). For perceptrons, there were 27 models in each the form diminutive and form non-diminutive group.

[7] This generalization excludes the Paradigm/Combined model at 50 epochs and a learning rate of .50. Table 3.4 shows that the F-score for this model has dropped to .330 , which is much smaller than the other paradigm models and likely due to overfitting to non-yer items.

The best decision tree for form and diminutive performed better than 22 of the corresponding perceptron models - the highest scoring perceptron, Form/Feature after 50 epochs with a learning rate of .30, scored 0.300, as seen in table 3.4. The form non-diminutive model performed better than 18 - the highest score was 0.522, and shared between the Form/Combined at learning rates of .10 and .30 and Form/Feature at learning rates of .10 and .50, all after 50 epochs. Building more models could further determine at what point the perceptrons begin overly rejecting tokens with yers and if there is a parameter setting that works for all, but the variability suggests that decision trees are able to competitively get results comparable with perceptrons.

3.7 Discussion

This study presented a comparison between CART decision trees and perceptrons for learning patterns of yers in Polish. Perceptrons can perform better than decision trees given correct parameter tuning, but they can be unstable in that they overfit after a point. The J48 decision tree algorithm used here may perform marginally worse than the best model in each model category but nonetheless stays competitive. As a general note, the decision tree models also took significantly less time to run, with all models taking less than 5 seconds to build.[8] The perceptron models, on the other hand, took anywhere from a few seconds with training time of 5 epochs to 45 minutes for training time of 50 epochs. The decision tree always correctly selected at least one item, while the perceptron had several instances where the model selected no items as having yers because it had learned no pattern at low training times and learning rates.

While the perceptrons were able to outperform the decision tree models, this raises the question of whether the percentages gained were worth the extra power used in training. These studies are, in the scheme of the computational world, inconsequential - nonetheless, in recent years, there is a higher focus on the ecological impact of large models that lack interpretability (Bender et al., 2021). Perceptron models are able to outperform decision tree models, yes, but the feature tuning required by these experiments to find such a model resulted in several that performed poorly at high training times, as shown in figure 3.2, or even that models with different training times performed best with different learning rates, as in table 3.6.

[8] These experiments were run on a Lenovo Thinkpad T480s, with 16 GB of RAM.

Conversely, this study also analyzed the amount of information necessary for computational systems to learn whether a yer exists in a word paradigm and whether a yer surfaces in a single given word. In particular, the inclusion of diminutives exposed a weakness in both groups of models, which were unable to learn the diminutive suffix due to the competition of many more non-yer environments that, due to the presentation of data, were indiscernable to the models from the yer environments.

Additionally, it examined the benefits of using a more interpretable system, like decision trees, compared to perceptrons for analytic purposes. While the perceptron may have slightly outperformed the decision tree, reading node weights makes corroboration with theory difficult. Decision trees, on the other hand, provide a more straightforward way of organizing data for analysis. In particular, this helps strike a balance in the amount of data need for generalizations. The decision trees had difficulty learning form yers until the presence of a paradigmatic yer was indicated as a variable, which suggests that if this variable were broken down into more specific, theory-driven pieces (for example, syllabic information, presence of a suffix), it could predict a yer without being directly told about one in the paradigm. The perceptron model, even when provided with this information, still failed to learn at low learning rate and training times.

This is particularly informative in the face of small data. Though the overall dataset is sizeable, there is a limited number of items that have yers, particularly when looking at the form level. Given this limitation, it was shown that the decision trees and perceptrons generally align in that they prioritize correctly selecting target items over expanding their generalization. However, the perceptrons appear much less likely to generalize to include new items at the expense of including non-target items. This interrogates how much data are needed by both techniques to retain these patterns in such a way that they is not overwhelmed by broader, unrelated patterns elsewhere in the data.

Moreover, this study provides an interesting contrast to the results of the phonotactic study in chapter 2. There, the best-performing model of Polish phonotactic acceptability was one that used full onset clusters instead of featural data, absent of frequency data. Here, regardless of whether the model was built using a decision tree or a neural network, the best performance came from models using fine featural detail. The models also paid closer attention to frequency, as certain patterns could not be learned when there was robust competition from a similar form that did not have a yer alternation. The difference between the phonotactic and morphological sets of experiments highlight the need to carefully determine what factors are at play during phonological experimentation and what their implications are for the larger representation.

3.8 Future work

Throughout, I have noted places where improvements can be made to the curation of the data set. Incorporating finer representations into Polish words, such as the syllabification in Rubach (2016), could add structure to the data set that is implied in the FWP model. Additional processing of the data could add features that allow other models to compete with the FWP models.

Throughout, I made reference to phonological theories that addressed where yers surfaced in extant words, as in Gouskova and Becker (2013) and Rubach (2016). However, neither approach fully addresses how a speaker may handle a novel word that may have an appropriate vowel in the appropriate position. They acknowledge that these words must be marked, but do not address how speakers decide whether there are yers in these new forms. Further investigation could examine whether the patterns predicted by these techniques are used by speakers when inflecting nonce words. Regarding other phonological patterns of Polish, such as [o]-raising, it has been argued that this process is no longer productive in some domains (Sanders, 2003), but conversely, that Polish speakers will replicate existing patterns in words that are 'similar' through means of analogy (Baranowski and Buckley, 2003). Finding a suitable domain for productivity is a question of balancing the features that speakers use in their representations. For example, in the case of German plurals, German-speaking children chose a smaller domain over which a regular rule could be more easily learned (Yang, 2016). Further investigation could determine where that boundary is for Polish speakers and yers, if one exists, and whether human and computational learning aligns in their solutions.

3.9 Conclusions

While perceptrons can achieve marginally higher accuracy given an appropriate training time and learning rate, decision trees appear more robust to change. In instances such as this learning problem, where the target items are such a small and narrow subset of data points, tuning the parameters of a perceptron results in slightly higher performance in exchange for much longer training times.

Here, I have examined benefits and detriments of both a CART decision tree and a perceptron for the same problem, outlining particular considerations that should be taken with each approach. I do find that perceptron models are able to outperform decision trees when given the same training data, so long as the model is tuned to appropriate parameter settings. However, finding which settings allow for the best-performing model can take up significantly more time and power that the decision trees, particularly in models with high training time. Nevertheless, both the perceptrons and decision trees both exhibited similar patterns in how they learned, or didn't learn, yer alternations given some data, suggesting they are forming similar representations. Notably, the diminutive pattern could not be learned due to outside competition. This is particularly apparent in the case of the FWP models, where informing the model whether there is an underlying yer significantly improves results. Determining what information is extrapolated from the paradigmatic variable will help determine what is being learned, what is in this representation, and whether the patterns are robust enough to be productive in new environments.

4: PERCEPTION

4.1 Introduction

Pierrehumbert (2016) asks what information a phonological representation must include. Certainly the environments that trigger phonological changes, aspects of the underlying and surface representations are needed, but Pierrehumbert draws attention to contextual information that is often, but not always, excluded. In this chapter, I use a psycholinguistic experiment to examine features that are similarly contextual and determine that, at the level of word access with regards to a prior priming task, speaker variation is *not* stored in the representation of that word. Other research on perceptual learning of phonemic categories shows that listeners *do* attend to this variation when it makes a phonetic features more apparent, so word access must operate with a different representation that excludes this information.

Pierrehumbert (2016) discusses episodic memory in the phonological space and outlines four factors that influence phonological representations and production at given contexts. She lists them as ***Phrasal Context***, ***Frequency/Predictability***, ***Different Voices and Dialects***, and ***Indexical Information***. ***Phrasal Context*** can refer most obviously to sentence-level prosody (Gussenhoven, 2002), but also to language-specific focus cues (Vogel et al., 2016; Huang et al., 2018) and corresponding speaker effort (Heller and Pierrehumbert, 2011). ***Frequency/Predictability*** refers to the representation's place in memory and the time taken to access and produce it, with more frequent/highly predictable items having faster, less precise articulation than less predictable items (Gahl, 2008; Jurafsky et al., 2002). Pierrehumbert refers to episodic memory as the accumulation of 'phonetic memories', which can then be made concrete through theories of word formation based on frequency in the lexicon (Yang, 2016). As certain alternations become more prevalent in a lexicon, they are more easily acquired and can become productive rules. ***Different Voices and Dialects*** refer directly to the differing ways in which speakers of the same language pronounce words, but also how listeners are able to generalize these myriad pronunciations into a single representation (Kapadia et al., 2023). This can be the phonetic differences between speakers of different sexes, regional differences, but also the way in which representations can be shifted based on exposure to more information (Kraljic et al., 2008).

Of these factors, ***Indexical Information***, defined by Pierrehumbert (2016) as ‘information about the speaker, the social context, or the physical context’, can have the least obvious connection. This means that speakers will include particular social cues in their speech, whether wittingly or not (Labov, 1972), and similarly that listeners can identify social groups from this information (Weissler, 2022). However, with this definition established, it is nonetheless simple to see its connection with the other four factors, particularly ***Different Voices and Dialects***. Phonetic differences based on sex characteristics (Diehl et al., 1996) and phonological differences from regional dialects (Labov, 1972, 2006) can easily be used to cue membership, in a way that goes beyond the phonetic differences between two speakers from the same social group. Additionally, as noted, some of these boundaries can be temporarily shifted based on recent perceptual input. This has been argued to provide category flexibility which allows listeners to incorporate dialectal information into their understanding (Sumner and Samuel, 2009; Kraljic and Samuel, 2005).

The flexibility based on speaker variation invites the larger question of what speaker-specific information is incorporated at what level of representational abstraction. Sumner and Samuel (2009) show that dialectal variation impacts word recognition, in that speakers of General American (GA) English are not primed by New York City (NYC) English words that exhibit r-dropping, but speakers of NYC English who exhibit r-dropping are nonetheless primed by the GA r-retaining variety. Meanwhile, Kraljic and Samuel (2005) demonstrate that the boundary between [s] and [ʃ] are shifted when primed through regimented exposure to recordings of a female voice but *not* of a male voice. This is attributed to higher pitch of the female voice compared to the male one, which makes the differences in sibilant frication more apparent.

In this chapter, I investigate the effect of multiple speakers on one such priming paradigm to relationship between ***Frequency/Predictability*** and ***Difference Voices and Dialects*** with ***Indexical Information*** by testing time taken to access sublexical units. Using a psycholinguistic experiment, I investigate whether a short-term priming effect generalizes between speakers - that is, if some word is known to have a priming effect on a second word when spoken by the same voice, will that effect hold if one item is spoken by a different voice? I examine this in particular through an *inhibitory* priming effect, one that makes it more difficult to reject a nonword (Sumner and Samuel, 2007). I begin by describing the language data used to construct priming paradigm and how each paradigm is constructed. I then explain how the study was conducted and its results. Finally, I compare these results to other studies of variation in speech to discuss varying stages of word processing and storage.

4.2 Previous work

The study of priming effects serves to clarify how words are stored and organized in the mind. 'Sublexical units' are sound items generally shorter than a whole word, which are used in language perception to decide on probable words based on matching contexts. This approach uses sublexical units as points during the processing timeline, a segment of sound that can then be interpreted as some root or affix - these subword units have been shown to have priming effect at the semantic level (Forster and Azuma, 2000; Rastle et al., 2004), sometimes even when in partial forms (Geary and Ussishkin, 2018). Moreover, a regular phonetic variation of a word will semantically prime a target item as successfully as its canonical form (Sumner and Samuel, 2005), though only the canonical form will lead to long-term activation.

Interest in how priming affects *nonwords* stems from the question of word recognition and selection, determining how the mind sorts through competing options to process a given sound (Marslen-Wilson et al., 1996; Marslen-Wilson, 1987; Luce and Pisoni, 1998). In particular, determining what features allow for these priming affects can inform how the mind interprets variation between speakers, what cues listeners attend to most readily (Mullenix and Pisoni, 1990). Sumner and Samuel (2005) investigated the effect of cues from multiple speakers on both short and long-term memory to find that regular and familiar variation patterns aid in processing, but only the canonical form has a long-lasting effect. They found, for example, that variant pronunciations of final [t] in GA English can still be processed as primes for the canonical [t]. That is, not only is 'music' semantically primed by [flut] ('flute'), variant [fluʔ] will also successfully prime, compared to [flus], which exhibits an irregular variation that does not map to the canonical form and thus has no priming effect.

Studying priming effects on *nonwords* can add to these discussions by determining what information is processed before the word can be rejected. Sumner and Samuel (2007) said that this access can be either facilitated or inhibited by priming speakers with exposure items that are similar except for a later sound. The 'network' of sublexical items stayed active for some time after initial exposure, making them more easily accessible for a short while. Listeners were told to decide whether a word was 'real' or 'fake' in two parts, where listening to items in the first part primed corresponding items in the second. Reaction times for this decision in the second stage found that being primed on a real word and then tested on a nonce word would result in a 20 ms delay in reaction time, wherein the listener's expectations of acceptability must be modified. When using these items to (inhibitively) prime nonwords, which have no inherent semantics, it suggests that there is an abstraction situated lower than the level of semantic connection but higher than individual phones. Furthermore Samuel and Dumay (2022) and Dumay et al. (2023) investigated how sleep affects this network activation and found that nonwords *can* be consolidated into the lexicon during sleep, suggesting there *is* an intermediary point at which these nonwords have an elevated status.

However, these experiments were conducted using a single speaker for both stages (Sumner and Samuel, 2007), or consistently used one speaker for priming and another for testing (Sumner and Samuel, 2009), which left the question of whether the delay found was speaker-specific or generalized among different speakers. This called into question how much phonetic information was stored in these sublexical representations preceding full words. Recall that Kraljic and Samuel (2005) found speaker-specific results for some experiments, when the speakers' differing phonetics affected when a phonetic boundary was moved based on the speaker's pitch making relevant features particularly salient. If the phonetic features of a speaker's voice are considered in perception, at what point are those details abstracted into a more general representation?

4.3 Experimental design

4.3.1 Language data

Unlike the studies reported in previous chapters, these experiments use English data. Using the stimulus list from Sumner and Samuel (2007) (specifically, for their experiment 1A), stimuli are monosyllabic words following English phonotactics, but are comprised of both real and nonce

Real Word	Final Cons	Vowel	Vowel
Prime	Change	Change	+ Cons
job	jop	jub	jup
[jɔb]	[jɔp]	[jʌb]	[jʌp]
solve	solb	silv	silb
[sɔlv]	[sɔlb]	[sɪlv]	[sɪlb]

Figure 4.1: Real-fake word paradigm used. ***Real Word Primes*** would prime ***Final Cons Change*** items, and ***Vowel Change*** items would prime ***Vowel + Cons*** items.

words, so long as they are well-formed monosyllables. The stimuli included simple and complex onsets and codas, but no vowel-initial or -final words.

Target tokens were sorted into three groups: 'real word primed', 'pseudoword primed', and 'unprimed'. Each word was part of a paradigm constituting a base 'real' word (***Real Word Prime***), a 'fake' word differing only in one feature of the final consonant (***Final Cons Change***), a 'fake' word differing only in the vowel (***Vowel Change***), and a 'fake' word differing in both the vowel and the same feature changed for ***Final Cons Change*** (***Vowel + Cons***). Items in ***Final Cons Change*** specifically can have changes in voicing ('job' and 'jop'), in place ('pluck' and 'plut'), or in manner ('crop' and 'croff'), but only in one of these features, which is then replicated for the ***Vowel + Cons*** token. This paradigm can be seen in figure 4.1. For the purpose of this experiment, the ***Real World Prime*** is meant to target the ***Final Cons Change*** token, and the ***Vowel Change*** token primes the ***Vowel + Voicing Change*** token, so the vowel easily distinguishes between a real prime and nonce prime.

Speakers were one female (late 20s and from the USA East Coast) and one male (early 70s and raised on the USA West Coast, though has been on Long Island for the past 30 years). Both were native General American (GA) English speakers, but their voices were distinctly discernible from one another.

4.3.2 Participants

Two hundred and forty participants were recruited through Prolific, a crowdsourcing platform, and took the experiment online. They were compensated $4.00 for the task, which took about 20 minutes to complete. Participants were restricted to ages 18-40 and, when prompted with a questionnaire, did not report a history of speech or hearing disorder (Woods et al., 2017). To

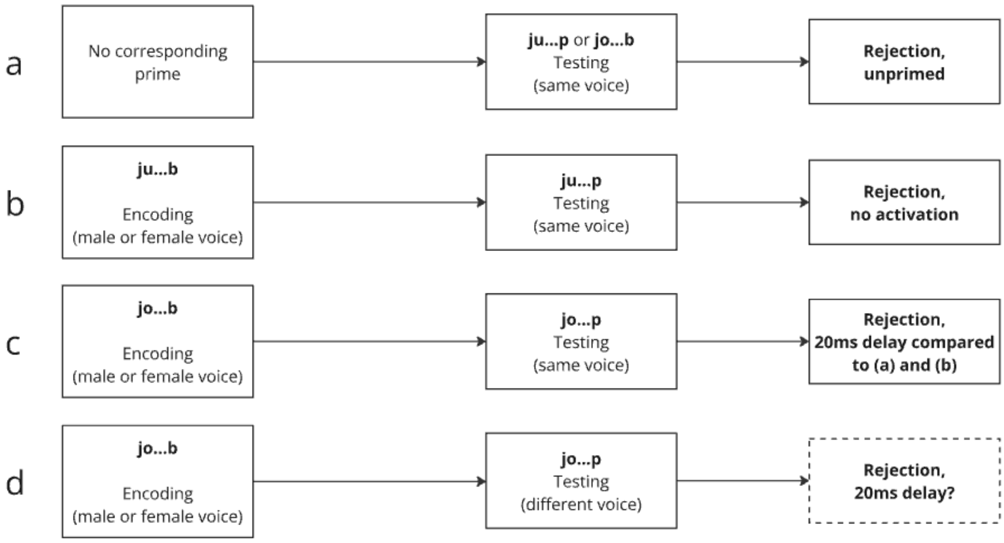

Figure 4.2: An outline of the experimental design for this experiment. (a) sets a control for an wholly unprimed item and (b) sets a control for a non-word that should not trigger sublexical activation. Sumner and Samuel (2007) show that (c), when primed and tested on the same voice, results in a delay, but did not investigate effects of different voices. The condition (d) is examined and tested here.

access the study, participants had to take a headphone check that asked them to rank the quietest of three sounds, in order to test the strength of their headphones and their own auditory abilities. Participants were native speakers of American English and residents of Connecticut, Delaware, Maryland, New Jersey, New York, Ohio, or Pennsylvania. This protocol was approved by the IRB of Stony Brook University.

4.3.3 Procedure

The experiment consisted of two parts, and encoding phrase that primed the participant and a test phase that measured their reaction time. In both halves, participants were presented a lexical decision task that asked them to decide whether the stimulus they had heard was a ‘real word’ or a ‘nonword’. Participants responded by pressing one of two buttons on their keyboard, corresponding with labeled answers. They were allowed a short break between the two tasks. Figure 4.2 shows this procedure with four distinct conditions. Conditions (a), (b), and (c) were studied in Sumner and Samuel (2007). This dissertation conducts (d).

In the encoding component, listeners completed the task for 18 ***Real Word Prime*** tokens and 18 ***Vowel Change*** tokens (which were paired with corresponding nonce words during the test

Encoding	Test	F(1,54)	Mean Squared Error	Significance
Male	Male	0.115	330.0083	0.736
	Female	2.574	11820.6750	0.114
Female	Male	1.120	3979.0083	0.295
	Female	5.356	12525.6333	**0.024**

Table 4.1: Single-factor ANOVA results for each encoding-test pair. The effect being tested is whether a target item was primed by a real word or a nonce word. Significant results ($p < .05$) are in bold.

component), from the word list used in Sumner and Samuel (2007). They also completed the task for 36 monosyllabic filler words, 36 disyllabic filler words, 36 monosyllabic 'fake' filler words, and 36 disyllabic 'fake' filler words. The monosyllabic filler words were selected to avoid using the same initial consonant-vowel (CV) or vowel-consonant (VC) sequence as in critical words; for disyllabic words, this holds within each syllable.

In the test component, listeners completed the task for the corresponding 18 ***Final Cons Change*** and 18 ***Vowel + Cons*** tokens task. Additionally, they completed the task for 18 new 'fake' words, to test against items that were unprimed (though controlled with regards to the test material). Of the six trial groups, group 1-4, 2-5, and 3-6 differed in whether these unprimed items were from the ***Final Cons Change*** list or the ***Vowel + Cons*** list, with all other items being the same. Fillers included 54 additional 'fake' words and 108 'real' words.

4.4 Results

A total of 266 participants were recruited, of which 26 were removed due to accuracy lower than 50%. This suggested that they had not paid sufficient attention to determine which words were real, so their results could not be reliable used with the assumption they had been successfully primed. Each participant was randomly assigned to one of 6 groups within each of 4 conditions, until each group had 10 complete participants.

Using a single-factor analysis of variance (ANOVA) as in Sumner and Samuel (2007), results indicate no effect of speaker voice. The main comparison of interest is whether the items primed by a real word behaved differently from an item primed by a nonce word, for each of the four combinations of prime and test speakers. Of the four conditions, only one (Female prime, Female test) gave a statistically significant result, $F(1,54) = 5.36, MSE = 12,525.63, p = 0.024$,

	Male Test	Female Test
Male Encoding - No Prime	1099 ms	1276 ms
Male Encoding - Nonse Prime	1122 ms	1268 ms
Male Encoding - Real Prime	1119 ms	1288 ms
Difference	-3 ms	20 ms
Female Encoding - No Prime	1169 ms	1319 ms
Female Encoding - Nonse Prime	1189 ms	1307 ms
Female Encoding - Real Prime	1201 ms	1327 ms
Difference	12 ms	20 ms

Table 4.2: Reaction time for test items that are primed by *nonse* and *real* words in the exposure task, plus the difference

shown in table 4.1.

Although the other three conditions were not statistically significant, raw averages nonetheless replicated the results found in Sumner and Samuel (2007), where nonce words primed by a real word had a 20ms slower reaction time than those primed by another nonce word. The unprimed condition, seen in table 4.2, sets the baseline for test items. Note that the reaction times for the ***Female Test*** conditions are approximately 200 ms longer than those of the ***Male Test*** conditions. To understand why, I measured the length of all stimuli recordings by each speaker. The female recordings averaged 1051ms, and the male recordings averaged 758ms. Reaction times were measured from the onset of the word, assuming that speakers had to listen to the entire recording to reliably judge whether it was a real word or nonce word. So, the difference in raw reaction times can be attributed, at least in part, to one speaker taking slightly longer to produce stimuli than the other speaker.

With this in mind, the target comparisons are the differences between the ***Unprimed*** experiment and ***Real Word Primed*** experiment compared to the differences between the ***Unprimed*** experiment and the ***Fake Word Primed*** experiment. Prior experiments showed a consistent inhibitory effect of 20ms for nonce primes, compared to real word primes (Sumner and Samuel, 2007). In table 4.2, we see the raw reaction times for all of the primed conditions, listed alongside the difference between the two conditions. Figure 4.3 shows these conditions grouped by both which voice was heard in testing and if that voice matched the encoding voice. It distinctly shows the delay of a 'real prime' over 'nonse prime' for both ***Female Test*** conditions, as well as the slight increase for the ***Female Encoding, Male Test*** condition.

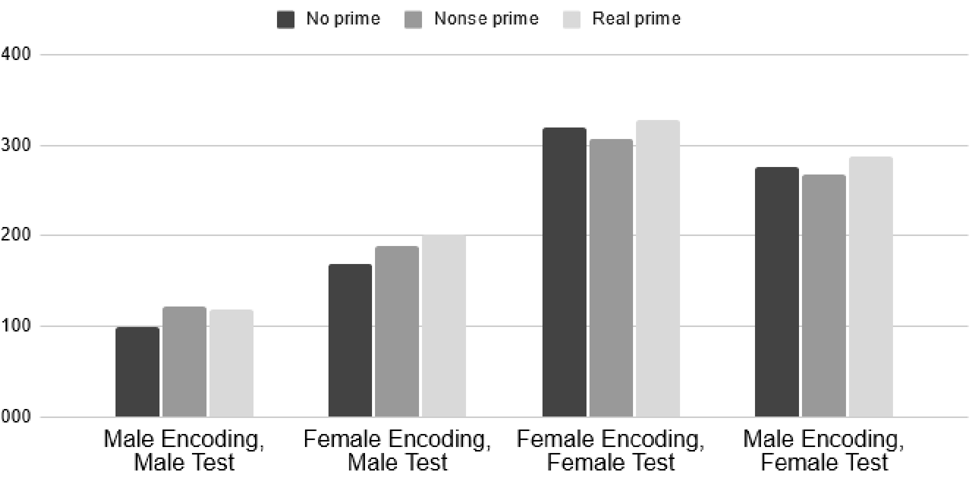

Figure 4.3: Reaction times per condition. Note that the Y-axis begins at 1000 milliseconds to highlight the difference between real and nonse reaction times.

While the raw reaction times exhibit the expected pattern, there are nonetheless complications. Firstly, participants who completed both the priming and test components listening to male recording failed to exhibit the expected reaction time difference. Secondly, recall that only the ***Female Encoding, Female Test*** condition was significant, even though both ***Female Test*** conditions showed the same difference in average reaction time. The ANOVA for ***Male Encoding, Female Test*** gives us $F(1, 54) = 2.574, MSE = 11,820.68, p = 0.114$, as seen in table 4.1. Third, the remaining condition - ***Female Encoding, Male Test*** - did show a reaction time difference in the expected direction, though it was not significant, with ANOVA values of $F(1, 54) = 1.12, MSE = 3979.01, p = 0.29$. The fact that this reaction time delay hold for ***Female Test*** conditions suggests that the effect generalizes between encoding-phrase voices, but should be replicated with additional voices for each condition.The combination of non-significant results and lack of reaction time delay in the ***Male Test*** condition suggest the issue may be with unconsidered features of the male voice or with the recordings.

4.5 Discussion

4.5.1 Sublexical access and speaker variation

This experiment was built on the foundation of Sumner and Samuel (2007), which ran the same experiment, but with a single female recorded for both priming and testing phases. The study reported here expanded on their results by testing whether having different speakers for the prime and test phases affected the inhibitory effect that real word primes had on nonce targets. Finding the same reaction time delay suggested that, while hearing the word in the priming phase did activate a mental representation of the word, this representation had abstracted away from phonetic features of the speaker. If, instead, it had not been abstracted away, we might have expected participants who were primed on female voices to have the expected delay in reaction time only when tested on female recordings, and vice versa for male recordings. Otherwise, if phonetic features of voice were stored at the word access level, the access could only be triggered by the same voice that had primed the listener.

The experiment done here focused on whether there was a change in reaction time when identifying non-words based on priming from the *same* voice or from a *different* voice. This assumes that, based on previous work of sublexical representations, the information stored when listening to priming material contributes to an abstract mental representation that is more readily accessible for a short period of time (Samuel and Dumay, 2022). Here, the experiment tests whether access to these representations is affected by 1) whether the test voice matches the priming voice, and 2) specific features of the priming voice. While the similar reaction times for the ***Male Exposure, Female Test*** and ***Female Exposure, Female Test*** conditions suggest that the priming voice has little effect, the differences between the ***Male Test*** and ***Female Test*** conditions are less clear. However, since phonetic variability is not stored in semantic priming tasks (Sumner and Samuel, 2005), it's likely that these unclear results are due to unconsidered features of the male voice or issues with the recordings.

Results from Kraljic and Samuel (2007) could inform this discrepancy and how to further investigate. When perturbing the boundary between [s]-[ʃ] and [t]-[d], they found that the first condition more reliably translated between speakers than the second. This was attributed to the characteristically higher frequency fricatives in the female voice compared to male voice (Jongman et al., 2000). Meanwhile, the voice onset time (VOT) that distinguishes [t] from [d] is less predictable by listeners (Allen and Miller, 2004; Allen et al., 2003). Though our experiment included a headphone check to determine whether participants were using headphones that could adequately represent the sounds, access to this sublexical representation could be affected differently depending on perceptual differences. Even if the phonetic features are not stored in such specific detail in the mental representation, a listener's ability (or inability) to perceive differences between sounds may affect word storage based on what phonetic features of a voice they can most easily attune to. This study used the word list from Sumner and Samuel (2007) (as well as in Samuel and Dumay (2022) later) for direct comparison, but further investigation into effect of *test* voice on sublexical access would require a more carefully balanced word list and closer attention to spectral differences.

4.5.2 Implications for abstract phonological representations

Returning to the factors of phonological representations listed by Pierrehumbert (2016), these results suggest that, solely in terms of sound processing, there are multiple levels at which these representations occur and this kind of word recognition occurs at a point that has abstracted away from phonetic features. It appears that at this level of word recognition, the most influential of these factors is ***Frequency/Predictability***, as the path in the map of sublexical units stays activated for some time to facilitate access during conversation. However, since different speakers did not affect the 20 ms reaction time difference between real-word priming and nonword priming, it can be assumed that ***Different Voices and Dialects*** plays less of a role at the level of sublexical abstraction.

It follows, then, that ***Indexical Information*** plays a similarly lessened role in the priming abstraction phase, since gender of the priming speaker did not have an effect. The role played at the test level is still unclear, but since prior research attributes perceptual differences to phonetic salience (Kraljic and Samuel, 2007), it's to be determined what effects are attributable to having multiple speaker identities, to phonetic differences between these speakers, and to perceived gender of the speakers.

It may yet be the case that there is an imbalance in priming caused by group-nonmembership. Inter-speaker flexibility can assist in accommodating dialectal variation. As shown by Sumner and Samuel (2009), the phonetic dialectal differences between General American (GA) speakers and New York City (NYC) speakers resulted in a pattern where both GA and NYC speakers were able to correctly perceive GA tokens, but GA speakers performed much worse on NYC tokens than the NYC speakers. Notably, the authors made the assertion that NYC speakers would have exposure to GA speakers (through media, schooling, etc.), but GA speakers were not guaranteed to have exposure to the NYC dialect. This exposure can affect the listener's own productive abilities, as evidenced through perceptual learning studies (Kraljic and Samuel, 2005, 2006, 2007), where recurrent exposure to ambiguous phonemes in a disambiguating context adjusted boundaries between minimally different phonemes. However, perceptual learning studies have shown that speakers are able to incorporate dialectal phonological patterns into their mental representation without necessarily producing it as a result (Kraljic et al., 2008).

Rather, it is likely the case that for the context-based environment provided by ***Indexical Information*** to be of use, the listener must first have incorporated the variety in ***Different Voices and Dialects*** into their mental representations. As discussed, Sumner and Samuel (2005) show that regular variation in pronunciation will still successfully prime target words, even if these phonetic differences are not stored in the long term representation. Sumner and Samuel (2009) conducted experiments to test listeners on their understanding of NYC English, considering three groups: speakers of NYC English, people who grew up with significant exposure to NYC English and understand it natively without producing phonological forms typical of NYC English, and people with no exposure to NYC English. They noted that the results of one participant were removed because "he complained that he could not do the task since some of the speakers 'could not speak English'". They also note that this participant had a particularly high error rate, categorizing most items as 'Pseudowords' in their lexical decision task. If we assume this participant was, in fact, performing the task to the best of their ability, this suggests particularly narrow phonological representations.

Sumner and Samuel do not elaborate further on this participant. However, I believe that the participant's high rejection rates are a manifestation of highly valuing the prestige GA dialect over the regional NYC dialect and careful attention to prescriptive rules of English. Pierrehumbert (2016) discusses ***Indexical Information*** in terms of speakers cuing what social group they belong to, but in this context, a participant may have instead *limited* their representation to signal the belief that this regional dialect is 'lesser', and that items they had incorrectly marked as 'pseudowords' did not conform to their more limited representation. Weissler (2022) notes the discrimination faced by speakers of African American English (AAE), even when they exhibit no morphosyntactic features typical of AAE in their speech, suggesting that there are phonological components of AAE that are recognizeable by nonspeakers of the dialect (Weissler and Brennan, 2020). As such, even when a speaker of AAE is not necessarily signaling group membership, the listener may have certain expectations, both conscious and unconscious of how speech should be produced. In this way, the factor of ***Indexical Information*** may differ from ***Different Voices and Dialects*** by speaking to the perception mechanism, instead of production.

The results of Sumner and Samuel (2009) and Weissler and Brennan (2020) speak to the boundaries in phonological representations, when an out-of-dialect speech signal cannot be incorporated into the representation. Kraljic et al. (2008) also finds that finding out-of-dialect forms acceptable in perception does not indicate they will be represented in production. Here, I have affirmed that, at the sublexical level, voice features has not been included in the abstraction, though it has been incorporated into the perceptual level of representation. This parallels how featural information seems to be excluded from representations in chapter 2 but are vital for representations in chapter 3. Here, when examining the ***Female Test*** condition, there is no difference in reaction time between encoded voices, and Sumner and Samuel (2009) showed there was a group of listeners that would perceive dialect-specific phonological patterns without difficulty, even if they did not produce those sound features themselves. Similarly adjusting the vocal features of the voice used in the priming task could determine how much phonological information is stored in a speaker's perception mechanism . For example, both male and female speakers of GA English can exhibit 'vocal fry', but female voices face more stigma when using it (Chen et al., 2002; Chao and Bursten, 2021). The female voice here did not use vocal fry, nor did it exhibit phonological features of AAE, so future work could determine if these added factors would change the listener's reaction.

4.6 Conclusion

This chapter has shown, using a lexical decision task, that there is a priming effect of real words on nonce words that abstracts away from speaker information. When given the task of rejecting nonce words that were primed by similar real words, prior research had shown a 20ms delay as the listener re-processed the word representation. Here, I replicated the task, scaling up from a single speaker to four combinations of a male or female voice of either part of the experiment. I found that, particularly for the condition tested on a female voice, this delay in reaction time was present regardless of whether stimuli were primed and encoded using a male or female voice.

This leaves an open question of the relationship between the perception mechanism, which takes phonetic distinctions into account, and the sublexical network tied to word access, which abstracts away from voice. Experiments that investigate the perception side of this relationship must take care to account for phonetic features that make some target sounds more salient. Similarly, since speakers may speak different dialectal variations that may not be recognized in the listener's phonological representation, additional work can determine what aspects of their dialects listeners attune to. I argue that perception must be considered in the phonological representation for this reason, as a rejection of dialectal speaker voice variation could negatively impact performance.

5: CONCLUSION

In this dissertation, I have addressed several learning problems, two computational and one psycholinguistic, with regards to phonological representations. In chapter 2, I looked at learnability of attested Polish word onsets and found that categorical and segmental models were better suited to an acceptability survey than frequency-informed featural models. In chapter 3, I used CART decision trees and perceptron neural networks to learn where alternating yers surfaced in Polish words depending on the morphophonological information available to each model. The perceptron models outperformed the decision trees, but not on all parameter settings, and both models exhibited similar struggles with regards to particular subsets of data - for example, generalizing over segmental information, particularly with regards to the diminutive suffix. In general, both decision trees and neural networks performed best when given featural information, which is in contrast to the phonotactic findings. In chapter 4, I examined effect of multiple speaker voices on nonce word rejection in a psycholinguistic experiment. Given different combinations of a female speaker and a male speaker as prime and test conditions, the experiment showed a similar reaction time delay in rejecting a nonce word similar to a *real* word they had previously heard. This delay was found for both male and female voice encodings, given the female test items. This showed that, after the initial exposure phase, the listener had created an abstract representation of the word absent of specific phonetic features that allowed later access regardless of speaker voice. This followed with a discussion of where gradient and featural distinctions were apparent in psycholinguistic problems, as in the distinction between the phonotactic and morphophonological experiment.

Together, these experiments showed that some levels of phonological processes rely heavily on specific featural information, while others abstract away from fine-grained details to create generalizations that could apply more broadly. These observations support closer analyses of how these processes operate on other languages and problems, but also how they interact with one another. I also highlight how results of computational and statistical linguistic experimentation are heavily shaped by how the experiments are conducted and the data used within. These results encourage careful attention to the amount of data, including phonetic details or word frequency, that are available to a model as it builds a representation, and to consider what implications those details have for human learning.

5.1 Future work

Throughout, I have addressed next steps for all of these questions. Regarding the question of Polish onset clusters, the most pressing question would be creating positive featural grammars to serve as direct correlates to the cluster and bigram models used. After that, expanding the scope of the question to include clusters that are more likely to be unfamiliar (i.e., all permutations of Polish consonants up to a certain cluster length) would examine how these generalize scale. Moreover, replicating these studies on other languages which allow complex onsets would be beneficial. As noted here, Polish yers result in complex clusters where they do not surface, which may result in different sonority patterns than clusters without underlying yers. Reporting results from languages that do not have additional concerns like these will better develop a theory of categorical phonotactics moving forward.

For the morphophonological models, there is an outstanding question of the information available in the FWP model through the paradigmatic variable. Decomposing each word form into more thorough phonological information - for example, including syllable structure - could improve model performance and create a more robust generalization. After determining a more precise environment for where yer alternations occur, this information can expand to the question of yer productivity. The environments predicted by these algorithms can be used as a basis for nonce word experiments with native speakers, to determine if the generalizations hold and if they have created a reproducible pattern or if speakers produce yer-like alternations in analogical environments, if at all.

Finally, the psycholinguistic experiment could be expanded to use more speakers, with a more carefully curated stimulus list. The results from tests with the female voice were much stronger than those with the male voice, which suggests either an issue with the male recordings used or that there is a secondary effect at play that blocks the priming effect. This follow up study would ensure a more carefully balanced word list so that phonological environment can be examined across voices to determine where differences occur.

5.2 Final conclusions

For applied machine learning, these results encourage closer attention to the information used to train models and how it may relate to the output. Specifically looking at yer alterations, there were several situations where a viable generalization was not learned due to both competition in the representation and the absence of information that could distinguish an alternating form from a non-alternating one. In comparison, the models that were told that there was *some* yer in the paradigm performed very highly. While giving the model access to yer information was in line with theories regarding the underlying representation of yers (Gouskova and Becker, 2013; Rubach, 2016; Scheer, 2018; Lightner, 1965), this approach would impact the generalizability of this model to new potential forms. If yer alternations are no longer productive - that is, if new words in the lexicon cannot exhibit this alternation, because the rule does not operate over a productive domain (Yang, 2016) - then these lexical representations may be desirable. However, Kraska-Szlenk (2007) notes use by children of the form [kɔmputra] in place of [kɔmputɛra], suggesting that children may at one point have acquired a yer-assignment generalization that is lost as they mature. This suggests that speakers do use morphophonological information to find a environment that regularly activates the alternation historically derived from yers.

Another major point is the reminder that the goal of the modeling problem can inform the data and structures needed for experimentation. Through, I have compared finer representations against broader ones, like whole clusters and symbols in place of feature combinations. However, there were additional considerations for why particular models performed poorly compared to others. In the case of modeling form yers, the sheer lack of word forms that had a surfaced yer made generalization difficult due to overwhelming competition from non-yer forms. For phonotactic models, the point made by Albright (2009) that negative grammars struggle to capture gradience in acceptability explains the poor performance of these models compared to ones built on positive grammars.

One care I took with this work was to use human-like data whenever possible, creating datasets from text corpora and whole morphological paradigms rather than a process in isolation, to see how the environment of a speech item affects how it is learned in a computational context. In settings that test live perception and production, the various levels of representation are more difficult to separate, but empirical studies take into consideration the effects of other processes when considering where in an operational pipeline the results sit. In my psycholinguistic experiment, I examined the very narrow question of how multiple speaker voices affects word access and rejection of nonwords. The positive results, in generalization across encoding voices for the female test condition, do support that sublexical access works at a more abstract level that excludes differences in voice. But the more unclear results, namely mixed results for the male test conditions, invite further discussion of how the test voice affected immediate processing. Perhaps certain features of characteristically female voices may have aided in the female test condition, or there was a subtle difference in dialect that nonetheless affected access to the activated network. The interaction of small processes on data curated to exhibit particular phenomena gives a richer understanding of how the larger phonological apparatus operates.

Here, I have argued for an approach to phonological study that favors using a wide variety of methods, with careful attention to how those methods impact the final results. By better understanding each part in detail, researchers are better able to understand how these pieces interact with one another. Computational methods are ideal for examining these small effects in isolation before bringing findings to experimental studies. This draws parallels with current approaches in artificial intelligence that argue for separate representations that interact with one another to formulate an explainable output (Swan et al., 2022). Human phonology works similarly, using different sets of variables at different levels of a representation such that their interactions operationalize language learning. With this in mind, careful attention to detail at differing levels of phonology enables stronger predictions for how these interactions play out based on the factors that come before. I have used this principle to closely examine two separate phonological question in similar terms - how much information is necessary for learning, and how detailed must that information be. I then bring these conclusions to a psycholinguistic question to break down where in the perception pipeline more detail is beneficial and where generalized abstractions facilitates processing. Through these observations and discussions, this dissertations helps us better examine the nature of phonological representations and the implications of modeling each level in particular ways.

A: APPENDICES

A.1 Polish onset clusters

t͡ʃ	1	fj	0.583	ktɕ	0	px	1	sxl	1
t͡ʃt͡ʃ	0.833	fk	0.583	kv	1	pxl	0.917	sxn	1
t͡ʃt͡s	0.667	fl	1	kw	1	pxw	0.917	sxr	0.667
t͡ʃk	1	fp	0.25	kɕ	0.333	pɕ	0.333	sxɲ	0.917
t͡ʃv	1	fr	1	kɲ	1	pɲ	1	sɕ	0.417
t͡ʃw	1	fs	0.417	kʃ	1	pʃ	0.833	sʃ	0.667
d͡ʒ	0.917	ft	0.583	kʃt	1	pʃt͡ʃ	1	t	1
d͡ʒd͡ʒ	0.917	fts	0.0833	kʒ	0.5	pʒ	0.583	tf	0.5
t͡s	1	ftɕ	0	l	1	r	1	tj	0.333
t͡sl	0.917	ftʃ	0.0833	lgn	0.917	rdz	1	tk	1
t͡sm	1	fx	0.5	lgɲ	1	rj	0.25	tl	1
t͡sn	1	fɕ	0	lj	0.417	rt	0.917	tm	0.75
t͡sv	1	fʃ	0.583	ln	0.917	rv	1	tn	1
t͡sw	1	g	1	lv	1	rʒ	0.917	tp	0.833
t͡sx	0.25	gd	1	lɕɲ	1	s	1	tr	0.917
t͡sʑ	0	gdʑ	0.5	lɲ	1	st͡s	0.917	trv	1
b	1	gjj	0	lʒ	0.917	sf	0.75	tsf	0.167
bj	0.583	gl	1	m	1	sk	1	tsm	0.333
bl	1	gm	1	mgl	1	skl	1	tsn	0.25
br	1	gn	1	mgw	1	skr	1	tv	1
brv	1	gr	1	mj	0.583	skw	0.917	tw	0.917
bw	1	gv	1	mkn	1	skʒ	0.5	tx	1
bz	1	gw	1	mkɲ	0.917	sl	0.917	tɕm	0.5
bʑ	0.417	gz	1	ml	0.917	sm	1	tɲ	1
bʒ	0.667	gɲ	1	mn	1	smr	1	tʃ	0.75
d	1	gʑ	0.5	mr	1	sn	1	tʃk	0.75
db	1	gʒ	0.667	mw	1	sp	1	tʃt	0.25
dj	0.583	gʒb	0.5	mx	1	spl	1	tʒ	0.5
dl	1	gʒm	0.583	mɟ	1	spr	1	tʒt͡s	0.0833
dm	1	j	1	mɲ	1	spʒ	0.5	tʒm	0.417
dn	1	jm	0.333	mʃ	1	sr	0.917	v	1
dr	1	k	1	mʒ	1	ss	1	vt͡ʃ	1
drg	0.917	kt͡s	0.75	n	1	st	1	vb	1
drv	1	kf	0.5	p	1	str	1	vd	1
drʒ	0.917	kjj	0.25	pj	0.583	sts	0.25	vdʑ	0.5
dv	1	kl	1	pl	0.917	stv	1	vg	1
dw	1	km	1	pn	1	stw	1	vj	0.667
dzb	1	kn	1	pr	1	stɕ	0	vk	1
dzv	0.917	kr	1	ps	1	stʃ	0.667	vkr	0.917
dɲ	1	krt	0.667	pstr	1	stʒ	0.5	vl	1
dʑ	0.583	krv	1	pt	1	sv	1	vm	1
dʑv	1	ks	1	pw	1	sw	1	vn	1
f	0.917	kt	1	pwt͡s	0.417	sx	1	vp	1

vpw	1	wz	1	zdj	0.833	ɕ	1	ʃp	1
vr	1	wʑ	0.417	zdr	1	ɕt͡s	0.333	ʃpr	1
vsk	1	wʒ	1	zdʑ	0.583	ɕf	0.333	ʃpʒ	0.333
vsp	1	x	1	zg	1	ɕl	1	ʃr	1
vst	1	xt͡s	1	zgj	0.75	ɕm	1	ʃt	1
vstr	0.917	xf	0.417	zgl	1	ɕp	0.917	ʃv	1
vstʒ	0.417	xj	0.583	zgr	1	ɕr	1	ʃw	1
vsx	1	xl	1	zgɲ	1	ɕv	1	ʑ	0.5
vt	1	xm	1	zgʒ	0.667	ɕɲ	1	ʑdʑb	1
vv	0.333	xr	1	zj	0.917	ɟ	1	ʑdʑbl	0.917
vw	1	xts	0.0833	zl	1	ɲ	1	ʑdʑbw	0.917
vz	1	xv	1	zm	1	ʃ	1	ʑl	1
vzg	1	xw	1	zmr	0.917	ʃt͡ʃ	1	ʑr	1
vzr	1	xʃ	0.667	zn	1	ʃt͡ʃv	1	ʒ	1
vzv	1	xʒ	0.583	zr	0.917	ʃt͡s	0.833	ʒb	1
vɕ	0.333	xʒt͡ʃ	0.417	zv	1	ʃf	0.667	ʒm	0.917
vɲ	1	xʒt͡s	0.25	zvr	1	ʃk	1	ʒn	1
vʃ	1	xʒt	0.333	zvw	1	ʃkl	1	ʒv	1
vʑ	0.333	z	1	zw	1	ʃkv	0.917	ʒw	0.917
vʒ	0.667	zb	1	zz	0.75	ʃkw	0.917	ʒɲ	1
w	1	zbl	1	zɲ	1	ʃl	1		
wb	1	zbr	1	zʃ	1	ʃm	1		
wg	1	zd	1	zʒ	0.917	ʃn	1		

BIBLIOGRAPHY

Adam Albright. Natural classes are not enough: Biased generalization in novel onset clusters. In *15th Manchester Phonology Meeting*, Manchester, UK, May 2007.

Adam Albright. Feature-based generalisation as a source of gradient acceptability. *Phonology*, 26(1):9–41, 2009. doi: 10.1017/S0952675709001705.

John Algeo. What consonant clusters are possible? *Word*, 29:3:206–244, 1978.

J. Sean Allen and Joanne L. Miller. Listener sensitivity to individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 115(6):3171–3183, 06 2004. ISSN 0001-4966. doi: 10.1121/1.1701898. URL https://doi.org/10.1121/1.1701898.

J. Sean Allen, Joanne L. Miller, and David DeSteno. Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, 113(1):544–552, 01 2003. ISSN 0001-4966. doi: 10.1121/1.1528172. URL https://doi.org/10.1121/1.1528172.

Esma Balkir, Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. Challenges in applying explainability methods to improve the fairness of NLP models. In Apurv Verma, Yada Pruksachatkun, Kai-Wei Chang, Aram Galstyan, Jwala Dhamala, and Yang Trista Cao, editors, *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*, pages 80–92, Seattle, U.S.A., July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.trustnlp-1.8. URL https://aclanthology.org/2022.trustnlp-1.8.

Maciej Baranowski and Eugene Buckley. Lexicalization and analogy in Polish o-raising, 2003.

Zsyzsanna Bárkányi. Gradient phonotactic acceptability: a case study from Slovak. *Acta Linguistica Hungarica*, 58(4):353–391, 2011.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Štefan Beňuš. Phonetic variation in Slovak yer and non-yer vowels. *Journal of Phonetics*, 40(3):535–549, 2012. ISSN 0095-4470. doi: https://doi.org/10.1016/j.wocn.2012.03.001. URL https://www.sciencedirect.com/science/article/pii/S0095447012000228.

Michael Biehl and Holm Schwarze. Learning by on-line gradient descent. *Journal of Physics A, Mathematical and General*, 28, 1995.

Thorsten Brants and Alex Franz. Web 1T 5-gram, 10 European Languages Version 1 lDC2009T25, 2009.

Leo Breiman, Jerome Friedman, R.A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman and Hall/CRC, 1984.

Eugene Buckley. Polish o-raising and phonological explanation. Washington DC, 2001. Annual Meeting of the Linguistic Society of America.

Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors. *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016*, Portorož, Slovenia, 2016. European Language Resources Association (ELRA), European Language Resources Association (ELRA). ISBN 978-2-9517408-9-1. URL `http://www.lrec-conf.org/proceedings/lrec2016/index.html`.

Jane Chandlee. Decision trees, entropy, and the contrastive feature hierarchy. *Proceedings of the Linguistic Society of America*, 8(1):5465, Apr. 2023. doi: `10.3765/plsa.v8i1.5465`. URL `https://journals.linguisticsociety.org/proceedings/index.php/PLSA/article/view/5465`.

Jane Chandlee, Remi Eyraud, Jeffrey Heinz, Adam Jardine, and Jonathan Rawski. Learning with partially ordered representations. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 91–101, Toronto, Canada, July 2019. Association for Computational Linguistics.

Monika Chao and Julia RS Bursten. Girl talk: Understanding negative reactions to female vocal fry. *Hypatia*, 36(1):42–59, 2021.

Yang Chen, Michael P Robb, and Harvey R Gilbert. Electroglottographic evaluation of gender and vowel effects during modal and vocal fry phonation. 2002.

Noam Chomsky and Morris Halle. *The sound pattern of English*. Harper & Row, 1968.

Robert Daland and Janet B. Pierrehumbert. Learning diphone-based segmentation. *Cognitive Science*, 35:119–155, 2011.

Robert Daland, Bruce Hayes, James White, Marc Garellek, Andrea Davis, and Ingrid Norrman. Explaining sonority projection effects. *Phonology*, 28:197–234, 2011.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. A survey of the state of explainable AI for natural language processing. In Kam-Fai Wong, Kevin Knight, and Hua Wu, editors, *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China, December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.aacl-main.46`.

Randy L Diehl, Björn Lindblom, Kathryn A Hoemeke, and Richard P Fahey. On explaining certain male-female differences in the phonetic realization of vowel categories. *Journal of phonetics*, 24(2):187–208, 1996.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017. URL `https://arxiv.org/abs/1702.08608`.

Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, second edition edition, 2001.

Nicolas Dumay, Sarah Kenway, Donghyun Kim, Efthymia Kapnoula, and Arthur Samuel. Do subphonemic mismatch effects only tell us about words, how they are learnt, and whether they need to sleep? 2023. Presented at the 64th Annual Meeting of the Psychonomics Society.

Emmanuel Dupoux, Kazuhiko Kakehi, Yuki Hirose, Christophe Pallier, and Jacques Mehler. Epenthetic vowels in Japanese: a perceptual illusion? *Journal of experimental psychology: human perception and performance*, 25, 1998.

Karthik Durvasula. Oh gradience, whence do you come? 2020. Invited plenary talk at the Annual Meeting on Phonology 2020.

David Eddington. A comparison of two tools for analyzing linguistic data: logistic regression and decision trees. *Italian Journal of Linguistics*, 22:265–286, 2010.

Peter Flach. *Machine Learning: the art and science of algorithms that make sense of data*. Cambridge University Press, first edition edition, 2012.

Kenneth I. Forster and Tamiko Azuma. Masked priming for prefixed words with bound stems: Does submit prime permit? *Language and Cognitive Processes*, 15(4-5):539–561, 2000. doi: 10.1080/016909600501196 98.

Eibe Frank, Mark A. Hall, and Ian H. Witten. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, fourth edition edition, 2016.

Ursula Franklin. *The real world of technology*. House of Anansi, 1999.

Susanne Gahl. Time and thyme are not homophones: the effect of lemma frequency on word duration in spontaneous speech. *Language*, 84:474–496, 2008.

Stephen I. Gallant. Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*, 1, 1990.

Jonathan A. Geary and Adam Ussishkin. Root-letter priming in Maltese visual word recognition. *The Mental Lexicon*, 13, 2018.

Kyle Gorman. *Generative Phonotactics*. PhD thesis, University of Pennsylvania, 2013.

Kyle Gorman. Pynini: a Python library for weighted finite-state grammar compilation. In *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata*, pages 75–80, 2016.

Kyle Gorman, Arya D. McCarthy, Ryan Cotterell, Ekaterina Vylomova, Miikka Silfverberg, and Magdalena Markowska. Weird inflects but OK: Making sense of morphological generation errors. In *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pages 140–151, Hong Kong, China, November 2019.

Maria Gouskova and Michael Becker. Nonce words show that Russian yer alternations are governed by the grammar. *Natural Language & Linguistic Theory*, 31(3):735–765, 2013.

Carlos Gussenhoven. Intonation and interpretation: phonetics and phonology. *Speech Prosody 2002, Interntional Conference*, 2002.

Edmund Gussman. *The Phonology of Polish* (*The Phonology of the World's Languages*. 2007.

Morris Halle. *Linguistic Theory and Psychological Reality*. MIT Press, 1978.

Ewa Haman, Bartłomiej Etenkowski, Magdalena Łuniewska, Joanna Szwabe, Ewa Dąbrowska, Marta Szreder, and Marek Łaziński. The Polish CDS corpus. Technical report, Talkbank, 2011.

Bruce Hayes. *Introductory Phonology*. John Wiley & Sons, 2011.

Bruce Hayes and Colin Wilson. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39:379–440, 2008.

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

Jeffrey Heinz, Chetan Rawal, and Herbert G. Tanner. Tier-based strictly local constraints for phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 58–64, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Jordana Heller and Janet B. Pierrehumbert. Word burstiness improves models of word reduction in spontaneous speech. *Proceedings of the 17th Annual Conference on Architecture and Mechanisms for Language Processing: poser session II*, pages 63–63, 2011. URL `https://amlap2011.files.wordpress.com/2011/08/postersession21.pdf`.

Yaqian Huang, Angeliki Athanasopoulou, and Irene Vogel. The effect of focus on creaky phonation in Mandarin Chinese tones. *University of Pennsylvania Working Papers in Linguistics*, 24, 2018.

Larry M. Hyman. How concrete is phonology? *Language*, 46:58–76, 1970.

Larry M. Hyman. *Phonology: theory and analysis*. 1975.

Gaja Jarosz. Polish yers and the finer structure of output-output correspondence. In *Annual meeting of the Berkeley Linguistics Society*, pages 181–192, 2005.

Gaja Jarosz. Implicational markedness and frequency in constraint-based computational models of phonological learning. *Journal of Child Language*, 37(3):565–606, 2010.

Gaja Jarosz. Defying the stimulus: Acquisition of complex onsets in Polish. *Phonology*, 34(2): 269–298, 2017.

Gaja Jarosz and Amanda Rysling. Sonority sequencing in Polish: the combined roles of prior bias & experience. 2016.

Gaja Jarosz, Shira Calamaro, and Jason Zentz. Input frequency and the acquisition of syllable structure in Polish. 2013.

Allard Jongman, Ratree Wayland, and Serena Wong. Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108(3):1252–1263, 09 2000. ISSN 0001-4966. doi: `10.1121/1.1288413`. URL `https://doi.org/10.1121/1.1288413`.

Daniel Jurafsky, Alan Bell, Michelle Gregory, and William D. Raymond. Probabilistic relations between words. *Frequency and the emergence of linguistic structure*, pages 229–254, 2001.

Daniel Jurafsky, Alan Bell, and Cynthia Girand. The role of the lemma in form variation. *Labratory Phonology*, 7:3–34, 2002.

Alexandra M. Kapadia, Jessica A. A. Tin, and Tyler K. Perrachione. Multiple sources of acoustic variation affect speech processing efficiencya). *The Journal of the Acoustical Society of America*, 153(1):209–209, 01 2023. ISSN 0001-4966. doi: 10.1121/10.0016611. URL https://doi.org/10.1121/10.0016611.

Maurice Kendall. A new measure of rank correlation. *Biometrika*, 30:81–89, 1938.

Christo Kirov and Ryan Cotterell. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. 6:651–665, 2019. URL https://www.aclweb.org/anthology/Q18-1045.pdf.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://www.aclweb.org/anthology/L18-1293.

Jordan Kodner. Computational approaches to morphology acquisition. In *Oxford Research Encyclopedia in Linguistics*. Oxford Univerity Press, 2022.

Kalina Kostyszyn and Jeffrey Heinz. Categorical account of gradient acceptability of word-initial Polish onsets. In Peter Jurgec, Liisa Duncan, Emily Elfner, Yoonjung Kang, Alexei Kochetov, Brittney K. O'Neill, Avery Ozburn, Keren Rice, Nathan Sanders, Jessamyn Schertz, Nate Shaftoe, and Lisa Sullivan, editors, *Proceedings of the 2021 Annual Meeting on Phonology*, Washington, DC, 2022. Linguistic Society of America. https://doi.org/10.3765/amp.v9i0.5317.

Tanya Kraljic and Arthur G. Samuel. Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51:141–178, 2005.

Tanya Kraljic and Arthur G. Samuel. Generalization in perceptual learning for speech. *Psychonomic Bulletin Review*, 13(2):262–268, 2006.

Tanya Kraljic and Arthur G. Samuel. Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56:1–15, 2007.

Tanya Kraljic, Susan E. Brennan, and Arthur G. Samuel. Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107(1):54–81, 2008. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2007.07.013. URL https://www.sciencedirect.com/science/article/pii/S0010027707002065.

Iwona Kraska-Szlenk. *Analogy: the Relation between Lexicon and Grammar*. 2007.

William Labov. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press, 1972.

William Labov. *The Social Stratification of English in New York City*. Cambridge University Press, 2006.

Theodore M. Lightner. *Segmental phonology of modern standard Russian*. PhD thesis, Massachusetts Institute of Technology, 1965.

Paul A. Luce and David B. Pisoni. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, 19:1–36, 1998.

William D. Marslen-Wilson. Functional parallelism in spoken word-recognition. *Cognition*, 25: 71–102, 1987.

William D. Marslen-Wilson, Helen E. Moss, and Stef van Halen. Perceptual distance and competition in lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 22:1376—-1392, 1996.

John P. Mullenix and David B. Pisoni. Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, 47:379–390, 1990.

Sarah Payne. A generalized algorithm for learning positive and negative grammars with unconventional string models. 2024.

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

Janet B. Pierrehumbert. Knowledge of variation. *Papers from 30th Regional Meeting of the Chicago Linguistic Society*, 2:232–256, 1994.

Janet B. Pierrehumbert. Stochastic phonology. *Glot International*, 5:195–207, 2001.

Janet B. Pierrehumbert. Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, 2:33–52, 2016.

J.R. Quinlan. *C4.5 : programs for machine learning*. Morgan Kaufmann Publishers, 1993.

Kathleen Rastle, Matthew H. Davis, and Boris New. The broth in my brother's brothel: Morpho-orthographic segmentation in visual word recognition. *Psychonomic Bulletin & Review*, 11: 1090–1098, 2004.

Jonathan Rawski. *Structure and Learning in Natural Language*. PhD thesis, Stony Brook University, 2021.

Jonathan Rawski and Jeffrey Heinz. "no free lunch in linguistics or machine learning: Response to Pater. *Language*, 95:125–139, 2019.

James Rogers and Geoffrey Pullum. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*, 20:329–342, 2011.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. On languages piecewise testable in the strict sense. In Christian Ebert, Gerhard Jäger, and Jens Michaelis, editors, *The Mathematics of Language*, volume 6149 of *Lecture Notes in Artifical Intelligence*, pages 255–265. Springer, 2010.

Jerzy Rubach. Polish yers: Representation and analysis. *Journal of Linguistics*, 1:1–46, 02 2016. doi: `10.1017/S0022226716000013`.

Zygmunt Saloni, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński, and Danuta Skowrońska. *Słownik gramatyczny języka polskiego*. Warsaw, 3rd edition, 2015. URL http://sgjp.pl.

Arthur Samuel and Nicholas Dumay. How active are sublexical and lexical representations 12 hours after they have been used to understand speech? 2022. Presented at the 63rd meeting of the Psychonomics Society.

Nathan Sanders. *Opacity and sound change in the Polish lexicon*. PhD thesis, University of California, Santa Cruz, 2003.

Edward Sapir. *Sound Patterns in Language*. Linguistic Society of America, 1925.

Tobias Scheer. Pn the difference between the lexicon and computation (regarding Slavic yers). *Linguistic Inquiry*, 50:197–218, 2018.

Robert J. Scholes. *Phonotactic grammaticality*. 1966.

Meghan Sumner and Arthur G. Samuel. Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language*, 52(3):322–338, 2005. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2004.11.004. URL https://www.sciencedirect.com/science/article/pii/S0749596X04001329.

Meghan Sumner and Arthur G. Samuel. Lexical inhibition and sublexical facilitation are surprisingly long lasting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33:769–790, 2007.

Meghan Sumner and Arthur G. Samuel. The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language*, 60(4):487–501, 2009. ISSN 0749-596X. doi: https://doi.org/10.1016/j.jml.2009.01.001. URL https://www.sciencedirect.com/science/article/pii/S0749596X09000096.

Jerry Swan, Eric Nivel, Neel Kant, Jules Hedges, Timothy Atkinson, and Bas Steunebrink. *The road to general intelligence*. Springer, 2022.

Jolanta Szpyra. Ghost segments in nonlinear phonology: Polish yers. *Language*, 68:277–312, 1992.

Irene Vogel, Angeliki Athanasopoulou, and Nadya Pincus. *Prominence, Contrast, and the Functional Load Hypothesis: An Acoustic Investigation*, page 123–167. Cambridge University Press, 2016.

Rachel Elizabeth Weissler. A meeting of the minds: Broadening horizons in the study of linguistic discrimination and social justice through sociolinguistic and psycholinguistic approaches. *Annual review of applied linguistics*, 42:137–143, 2022.

Rachel Elizabeth Weissler and Jonathan R. Brennan. How do listeners form grammatical expectations to african american language? *University of Pennsylvania Working Papers in Linguistics*, 2020.

Richard M. Weist and Katarzyna Witkowska-Stadnik. Basic relations in child language and the word order myth. *International Journal of Psychology*, 21:363–381, 1986.

Richard M. Weist, Hanna Wysocka, Katarzyna Witkowska-Stadnik, Ewa Buczowska, and Emilia Konieczna. The defective tense hypothesis: On the emergence of tense and aspect in child Polish. *Journal of Child Language*, 11:347–384, 1984.

Marcin Woliński and Witold Kieraś. The on-line version of Grammatical Dictionary of Polish. pages 2589–2594.

Kevin Woods, Max Siegel, James Traer, and Josh McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, Psychophysics*, 79, 07 2017. doi: 10.3758/s13414-017-1361-2.

Charles Yang. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language*. 2016.

Jennifer Yearley. Jer vowels in russian. *Papers in optimality theory*, pages 533–571, 1995.

Chih-Kuan Yeh, Been Kim, and Pradeep Ravikumar. Human-centered concept explanations for neural networks. *Neuro-Symbolic Artificial Intelligence: The State of the Art*, 342, 2022.

Paulina Zydorowicz and Paula Orzechowska. The study of Polish phonotactics: Measures of phonotactic preferability. *Studies in Polish Linguistics*, 12:97–121, 2017.