Towards an Algorithmic Account of Phonological Rules and Representations

by

Caleb A. Belth

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Computer Science and Engineering)
in the University of Michigan
2023

Doctoral Committee:

Professor Danai Koutra, Co-Chair
Professor Andries Coetzee, Co-Chair
Professor Richard Lewis
Professor Lu Wang
Professor Charles Yang, University of Pennsylvania

Caleb A. Belth
cbelth@umich.edu
ORCID iD: 0000-0002-6256-3381

ACKNOWLEDGEMENTS

When asked how I got interested in the study of language from a computer science background, I often credit the more or less parallel events of learning about the computational theory of mind in Dan Kelly's philosophy of mind class at Purdue, and the coincidence of my favorite study spot being the HSSE library, right next to the linguistics bookshelves. But, as is always the case in trying to establish "first causes" in history, an infinite number of alternative origin points could also have been chosen. For example, I could point to my parents choosing a grammar textbook that so drilled into me the process of diagramming a sentence that I couldn't help but see its algorithmic character. Or I might choose as the starting point a google scholar search that turned up Charles Yang's paper (with others) on principles of computational efficiency in language, or credit the first graduate linguistics class being Andries Coetzee's phonology course that directed me to the domain of sound patterns in particular, and away from the perils of syntax. I might point to the yearly laughs elicited by Chad Gadya's recursive "Then came the butcher and killed the ox that drank the water that quenched the fire that burned the stick..." giving the distinct sensation of voices rolling down a hill. The countless options for labeling as the paramount moment in a narrative of my academic journey reveal the countless numbers of people who have supported and guided me in that journey. For the next paragraphs, I'd like to highlight a few of those people.

I'll start by thanking my advisors, Danai Koutra and Andries Coetzee. When I entered graduate school, I thought that I wanted to do natural language processing but, lacking much experience in NLP research, I didn't get in to any NLP programs. Danai always gave me autonomy in my research—from day one—even when she was supporting me with her own grant money. This has made my PhD experience an intellectual joy. She was also instrumental in providing top-notch feedback on my NSF fellowship application, which ultimately provided the financial support needed to pursue research goals more in congruence with my subjective sense of interest. Among these things, she also did many other big things like teach me the value of a good figure, and of minimizing text on powerpoint slides; she thought me countless LaTeX tricks, and how to pronounce a good chunk of the Greek alphabet, which turned out to be of no small use when it came time to learn the International Phonetic Alphabet.

Andries introduced me to phonology. Much more importantly, he never let me feel like an imposter in linguistics. He has been a fierce advocate for me, allowing me to take his graduate

level phonology course with no phonological training, encouraging me to turn my term paper into a journal article, advocating for me to teach linguistics classes, and on.

Despite being at a different institution, Charles Yang effectively adopted me as his own student, and always believed in me, provided insight and inspiration, and influenced my thinking about the implications of my work.

Rick Lewis and Andries both provided valuable guidance in human-subjects experimental methodology. Meetings with Lu Wang were important to my thinking about the relationship between my work and neural networks.

My whole committee has been incredibly supportive, especially with flexibility and kindness during my job search, which has allowed my stress to stay at tolerable levels.

Beyond my committee, I want to thank Jeff Heinz, Jane Chandlee, and Kyle Gorman for providing valuable feedback on my work. I am honored that Charles Yang, Jordan Kodner, and Sarah Payne adopted me as a Penn Pal.

Turning to friends in Ann Arbor, the brightest moments of the pandemic came on weekend hangouts with Ashkan and Alican, who formed my pandemic bubble. It was a certain video on paranormal activity that led to unquestionably the hardest laughing of my PhD. If you know you know.

The so-called "Thursday Dinner crew" has been there all along. Eli, Fahad, Kevin, Sarah, Trevor, and Won have all been there for me through the emotional toil that goes with being a PhD student, and even more so with being a human being. But we've also shared a vast chunk of the memories I'll take with me from my PhD.

I'm very grateful for meeting Santiago, Ashkan, Laura, and Oana in early PhD courses, and for the shared memories that ensued. I am also thankful for friendships with fellow GEMS lab members, Tara, Mark, Marlena, Jiong, Puja, Yujun, and Di.

The last year would not have been the same without my adventure partner, Brittany, who has traveled the country with me, and made long working weekends to nevertheless seem like they were a vacation. We've scoured second-hand bookstores together. Your spirit constantly reminds me that each experience is not preparation for future living, but an act of living itself.

I also want to thank my family for providing a space for me to rest beyond the academic world. You all have always been there to celebrate and support me through this journey. I am especially grateful for my big sister Rachel, who has shown me kindness, understanding, and guidance in ways that have been enormously significant to me, and my Grandpa Joe who embodies integrity in an academic life

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES viii

LIST OF TABLES x

LIST OF APPENDICES xiii

ABSTRACT xiv

CHAPTER

1 Introduction 1

1.1 Contributions and Dissertation Structure 5
1.1.1 Change of Representation When and Only When Needed 5
1.1.2 Expansion of Attention When and Only When Needed 7
1.1.3 Experimental Validation of Model Predictions 8

2 Background 10

2.1 Morphophonological Alternations 10
2.1.1 Productivity, Underlying Forms, and History 11
2.1.2 Non Adjacency 13
2.2 Sparsity in Language 14
2.2.1 The Frequency Distribution of Language Requires Generalization 14
2.2.2 The Indefinite Size of Language Requires Generalization 15
2.2.3 Paradigm (Un)saturation Requires Generalization 15
2.2.4 Summary 17
2.3 Studies Pertaining to Lexical Representations 17
2.3.1 Lexical Studies 17
2.3.2 Experimental Studies 18
2.4 Experimental Studies of Sequence Learning 19
2.4.1 Domain-Independence of Adjacent Dependency Tracking 21
2.4.2 The Developmental Trajectory of Sensitivity to Adjacent and Non-Adjacent Dependencies 21
2.4.3 Adjacency as Default 22
2.4.4 Summary: A Proclivity for Adjacency 23

2.5 The Tolerance Principle . . . 23
2.5.1 The Role of the Tolerance Principle . . . 24

3 An Algorithmic Account of Phonological Tiers . . . 25

3.1 Introduction . . . 25
3.2 Model Description . . . 29
3.2.1 The Structure of Generalizations . . . 29
3.2.2 Learning . . . 33
3.2.3 Strict Locality as a Special Case . . . 39
3.3 Prior Models . . . 39
3.3.1 Statistical Models . . . 40
3.3.2 Formal-Language-Theoretic Approaches . . . 41
3.3.3 Neural Network Models . . . 42
3.4 Comparison to Human Behavior . . . 42
3.4.1 Model Behavior on Finley (2011) . . . 42
3.4.2 Model Behavior on McMullin and Hansson (2019) . . . 47
3.5 Learning Natural Language Alternations . . . 51
3.5.1 Turkish Vowel Harmony . . . 51
3.5.2 Finnish Vowel Harmony . . . 53
3.5.3 Latin Liquid Dissimilation . . . 55
3.6 Discussion . . . 57
3.6.1 D2L vs. Other Models . . . 58
3.6.2 Future Directions . . . 61

4 Towards an Algorithmic Account of Underlying Forms . . . 62

4.1 Introduction . . . 62
4.2 Model . . . 64
4.2.1 Model Input . . . 64
4.2.2 Model Output . . . 65
4.2.3 Model Description . . . 66
4.2.4 When is Abstraction Needed? . . . 67
4.2.5 Constructing Abstract URs . . . 70
4.2.6 Learning Alternations . . . 70
4.3 Turkish Case Study . . . 71
4.3.1 Turkish . . . 71
4.3.2 Setup and Data . . . 73
4.3.3 Suffixes: Abstract and Concrete . . . 74
4.3.4 Quantitative Evaluation . . . 76
4.4 Dutch Case Study . . . 78
4.4.1 Dutch . . . 78
4.4.2 Setup and Data . . . 81
4.4.3 Generalization Without Abstraction . . . 82
4.4.4 Quantitative Evaluation . . . 85
4.5 Prior Work . . . 86
4.6 Conclusion . . . 87

4.6.1 Limitations and Future Directions . . . 88

5 An Algorithmic Account of Phonological Rules . . . 89

5.1 Introduction . . . 89
5.1.1 Locality . . . 91
5.1.2 The Nature of the Learning Task . . . 92
5.1.3 Locality and Identity as Principles of Computational Efficiency . . . 94
5.2 Model: PLP . . . 94
5.2.1 The Input . . . 95
5.2.2 Constructing Generalizations . . . 96
5.2.3 Encoding Generalizations in a Grammar . . . 101
5.2.4 Updating Incrementally . . . 106
5.3 Prior Models . . . 107
5.3.1 Constraint-Based Models . . . 107
5.3.2 Rule-Based, Neural Network, and Linear Discriminative Models . . . 108
5.3.3 Formal-Language-Theoretic Models . . . 109
5.4 Evaluating the Model . . . 110
5.4.1 Model Comparisons . . . 111
5.4.2 Comparison to Humans' Preference for Locality . . . 112
5.4.3 Learning German Devoicing . . . 115
5.4.4 Learning a Multi-Process Grammar . . . 120
5.4.5 Learning Tswana's Post-Nasal Devoicing . . . 123
5.5 Discussion . . . 125
5.5.1 Future Directions . . . 125

6 Experimental Evaluation . . . 127

6.1 Introduction . . . 127
6.2 Methodology . . . 129
6.2.1 The Artificial Language . . . 129
6.2.2 Participants . . . 131
6.2.3 Stimuli . . . 131
6.2.4 Experimental Design . . . 132
6.2.5 Hypotheses . . . 134
6.3 Evaluation . . . 137
6.3.1 Effective Learning and Default Affix . . . 137
6.3.2 Hypothesis Evaluation . . . 141
6.4 Model Comparison . . . 144
6.4.1 Setup . . . 144
6.4.2 Results . . . 145
6.5 Disucssion . . . 149
6.5.1 Algorithms and Formal Language Theory . . . 150
6.5.2 Future Directions . . . 151

7 Conclusion and Discussion . . . 152

7.1 Main Results and Contributions . . . 152

7.1.1 Adjacency First . . . 152
7.1.2 Change in Representation as the Consequence of Adjacency First . . . 153
7.1.3 Productivity and Lack Thereof . . . 153
7.1.4 The Typologically Rare . . . 153
7.2 A Critical Comparison with Neural Networks . . . 154
7.2.1 Neural Networks as Computational Theories of Mind . . . 155
7.2.2 Learning Algorithms and Developmental Predictions . . . 156
7.2.3 Accuracy on Small Data . . . 157
7.2.4 Non-Human-Like Generalization Beyond Training Distribution . . . 158
7.2.5 Interpretability . . . 159
7.2.6 Potential Contributions to Sample-Efficient NLP and ASR . . . 159
7.3 Limitations and Future Directions . . . 162
7.3.1 Putting Pieces Together . . . 162
7.3.2 Lower Levels of Representation . . . 163
7.3.3 Variation . . . 163
7.3.4 Typology . . . 164
7.3.5 Rule Interaction . . . 165
7.3.6 Tonal Phonology . . . 167
7.3.7 Connections to Syntax . . . 167

APPENDICES . . . 169

BIBLIOGRAPHY . . . 181

LIST OF FIGURES

FIGURE

2.1 The log frequency of child-directed word types in Brown (1973)'s acquisition corpus approximately follows the famous Zipfian distribution, in which a few words occur very often and most words occur only a few times. 15
2.2 The rate of new child-directed word types in Brown (1973)'s acquisition corpus decreases as the number of tokens increases, but never approaches an asymptote. 16
2.3 The developmental trajectory of learners' ability to track adjacent and non-adjacent dependencies. Infants as young as 8mo old show evidence of being able to track adjacent dependencies; this ability persists into adulthood. In contrast, do not show evidence of tracking non-adjacent dependencies until around 15mo. 22

3.1 A visualization of our proposed algorithm running on a toy example of sibilant harmony. 28

4.1 Our proposed model's accuracy (averaged over 5 simulations) at predicting novel surface forms. The *x*-axis shows the growth of the learner's lexicon (i.e., the training size). The gray, dashed line marks the 0.9 accuracy point. 77
4.2 Morphological Inflection Rules 83
4.3 Accuracy generalizing to held-out test words. 86

5.1 **(a)**: The width of PLP's search is expanded (upward arrows) when and only when an adequate generalization cannot be found in virtue of a less-wide context. **(b)**-**(c)**: An example of PLP's search: the first generalization (101) fails because it makes too many wrong predictions, but the second (103a) allows the /z/ → [s] instances to be isolated. . 99
5.2 PLP best matches the locality results of Baer-Henney and van de Vijver (2012), where participants much more easily learned languages with local generalizations (languages 1-2) than a non-local generalization (language 3). In contrast, MGL fails to mirror these results, learning all generalizations equally well. Grammars constructed by ranking a provided constraint set also fail to match the results: if provided with phonetically-motivated constraints, only the first generalization can be learned, and if provided with all or language-relevant constraints, all generalizations are learned equally well. 116
5.3 PLP's accuracy on the plural and past tense nonce words from Berko (1958) as training progressed. The black dashed line denotes plurals that should take [-z] or [-s] and the gray dashed lines those that should take [-əz]. The dotted lines represent the analogues for past-tense. The fact that [z]/[s] accuracy converges before [-əz] and [d]/[t] before [-əd] matches Berko (1958)'s finding that children learn [-z]/[-s] and [-d]/[-t] before [-əz] and [-əd]. 123

6.1 Example images for stimuli. 133
6.2 Accuracy on Train-Like Test Items. 138
6.3 The distributions of which affix form was chosen. 139
6.4 The distributions of which affix form was chosen for novel test items. 143
6.5 Model accuracies on Train-Like Test Items, compared to humans. 145
6.6 The distributions of which affix form was chosen for novel test items. 148

LIST OF TABLES

TABLE

1.1 Summary of Contributions 9

2.1 A hypothetical state of a child's mental lexicon for some English nouns. The child only knows both the SG and PL forms for some words. 17

3.1 Artificial Language Results - Finley (2011) Experiment 1 45
3.2 Artificial Language Results - Finley (2011) Experiment 2 46
3.3 Results from McMullin and Hansson (2019)'s Experiment 1a (assimilation) and 2a (dissimilation), where training instances involved liquids interacting across intervening vowels. D2L matches human behavior in all cases. 49
3.4 Results from McMullin and Hansson (2019)'s. Experiment 1b (assimilation) and 2b (dissimilation), where training instances involved liquids interacting across intervening vowels. D2L matches human behavior in all cases. 49
3.5 Accuracy of models on held-out test words, when learning Turkish vowel harmony. . . 53
3.6 Accuracy of models on held-out test words, when learning Finnish vowel harmony. . . 55
3.7 Accuracy of models on held-out test words after learning Latin liquid dissimilation. . . 56

4.1 An example Turkish input consisting of morphologically-analyzed surface forms. . . . 65
4.2 When the first three words from Tab. 4.1 enter the lexicon, the stems and plural affix are all stored concretely (left two columns). The plural form of the 'ice' and 'girl' stems are predictably decomposable into their concrete stems and the PL affix (denoted with the boldface concatenation), so those forms need not be stored in the lexicon. However, with /-lɑr/ as the UR of the plural, the plural form of 'hand' cannot be so decomposed, so it is instead lexicalized. 67
4.3 The left two columns contain morphemes—meaning and form (UR); the right three columns contain word forms. Boldface denotes word forms that can be predictably decomposed into concrete underlying forms, while '/-/' notation denotes word forms that must be lexicalized. The '??' denotes word forms that are unknown. Once all nine words from Tab. 4.1 enter the lexicon, most forms (6 of 9) cannot be predictably decomposed into concrete underlying forms, so the model constructs abstract URs, as described in § 4.2.5. 68
4.4 Top 10 most frequent affixes in a random, frequency-weighted sample of 1K words from the CHILDES dataset, and the URs that our model learned. See http://coltekin.net/cagri/trmorph/trmorph-manual.pdf for a description of affix names. 75

4.5 The underlying forms of the plural and diminutive suffixes, as well as example roots 'horse' (which alternates) and 'book' (which does not alternate) after running on the model on the 887 Dutch nouns from the CHILDES dataset. 82
4.6 Both humans and our model perform much better at identifying the singular of a nonce plural when the noun does not alternate than when it does. In comparison, a trigram language model does not reflect this asymmetry. 84
4.7 The logic of model's functioning, and how high generalization accuracy is achievable both when abstraction is needed and when it is not. 88

5.1 Model accuracies (with standard deviations) on held-out test data at different training vocabulary sizes. PLP readily learns an accurate generalization for German final-obstruent devoicing. 119
5.2 Analysis of the types of errors each of the models that learn an accurate grammar make in the process. Because it only adds generalizations to the grammar when necessitated by surface-alternation, PLP produces no unmotivated errors. 119
5.3 Model accuracies (with standard deviations) on held-out test data at different training vocabulary sizes. PLP readily learns an accurate grammar for the English processes in (132). 121
5.4 PLP learns precisely the set of processes active in its experience. This provides a straight-forward account of how productive phonological processes can be learned even if they operate against apparent phonetic motivation, like devoicing in Tswana following nasals (Coetzee and Pretorius, 2010). With PLP, the unmotivated constraint *NÇ need not be assumed universal. 124

6.1 Segment Inventory 131
6.2 Training Data 131
6.3 Test Data 132
6.4 The affix predicted by LOCAL and CONSERVATIVE for Experimental Group participants when presented with Novel Test instances. Cells with '??' denote a prediction that the Experimental Group should show no preference for one affix over the other. 136
6.5 Fixed-effects component of the mixed model fit to the responses to train-like test items. 140
6.6 Pairwise Z-tests for the Estimated Marginal Means between [+cons,−voi][+vowel,−back] and [+cons,+voi][+vowel,+back] for each Group. 141
6.7 Pairwise Z-tests for the Estimated Marginal Means between Control and Experimental groups for each Type of train-like test item. 141
6.8 Fixed-effects component of the mixed model fit to responses to novel test items. 143
6.9 Pairwise Z-tests for the Estimated Marginal Means between Control and Experimental groups for [+cons,+voi][+vowel,+back] and [+cons,−voi][+vowel,−back]. 144
6.10 The difference between humans' choices and models' choices. The first two columns show the absolute difference in average rate of choosing the [-f] form for [+cons,−voi][+vowel,+back] items and the absolute difference in average rate of choosing the [-ʃ] form for [+cons,+voi][+vowel,−back] items. The third column summarizes these for each model by averaging the first two columns. 148

7.1 Summary of neural network results in comparison to our proposed models. 158

A.1 Features for comparison to Finley (2011). 171
A.2 Features for comparison to McMullin and Hansson (2019). 171
A.3 Features for Turkish. 172
A.4 Features for Finnish. 173
A.5 Features for Latin. The consonant features are taken from Cser (2010). 174

C.1 100 Training Pairs. The Experimental Group is presented the pairs, the Control Group is presented with only the stems. 179
C.2 52 Test Pairs—the same for the Experimental and Control Groups. 180

LIST OF APPENDICES

A Chapter 4 Appendix **169**

B Chapter 3 Appendix **175**

C Chapter 6 Appendix **179**

ABSTRACT

The development of computer science in the middle of the twentieth century provided a valuable tool for the study of language as a cognitive system, by allowing linguistic theories to be stated in computational terms. The resulting theories have traditionally placed emphasis on describing the space of possible human languages, and viewed this delineated space as antecedent to a theory of how such a language might be learned from linguistic data. In the domain of phonology—the study of the structure of linguistic sound—this dissertation takes steps approaching the problem from the opposite direction, by framing the problem as that of identifying the learning procedure(s) by which humans construct a language in response to linguistic exposure. The object of study is shifted from the investigation of how a learner will discover a supposed target grammar, to the investigation of the ontogenetic process by which humans develop computational, phonological systems.

The proposed algorithmic approach identifies independently-established psychological mechanisms available to a learner, and then uses these as the components of a hypothesized learning procedure. The dissertation includes an algorithmic account of how graph-based representations of words, which render long-distance dependencies as local in that graph structure and are known as phonological *tiers*, arise naturally from a learning algorithm sensitive to only adjacent dependencies. The dissertation also proposes an algorithmic account of when abstract representations of morphemes are needed for effective generalization to unseen words in the face of the sparsity of linguistic input, and how rules can be constructed to map between these abstract representations and their concrete realizations. Stated in explicit, computational terms, the proposed learning system is evaluated on realistic natural language data, and makes precise, testable predictions. The learner constructs accurate linguistic generalizations from naturalistic data: across languages evaluated, the learner achieves, on average, 0.96 accuracy on held-out test words, and never lower than 0.92. These results are achieved with training data of no more than a thousand words. Moreover, the models' predictions are consistently borne out in developmental predictions and experimental settings, including a novel experiment carried out to directly test this model.

When compared to a prominent alternative learning-based model—neural networks—the proposed model achieves higher accuracy, while producing comparatively interpretable outputs, and—critically—providing an intelligible algorithm, which brings greater understanding to the mechanisms underlying phonological development.

CHAPTER 1

Introduction

There is nothing to which growth is relative save more growth.
– John Dewey (1916)

Since at least Ferdinand de Saussure (1916)'s *signifiant* and *signifié*, words have often been thought of as a mapping between *form* (sound or sign) and *meaning*. A word's mental representation could be concrete, homomorphically reflecting the physical properties of its form as a sound or sign. But linguists have often been motivated to hypothesize various abstract representations, which do not directly reflect the physical realization of the word form and thus require a computational system to derive concrete "surface" representations from abstract "underlying" representations and vice versa. Abstract representations can allow the distributional properties of word parts (morphemes) to be predicted and productively extended to novel words, or capture dependencies in a more appropriate data structure (e.g., *autosegmental tiers* and *metrical grids*) than flat sequences. If word forms are stored concretely in a mental *lexicon*, then they can be used in production and comprehension of linguistic expressions via direct access. On the other hand, if the representations involve some degree of abstraction, then a *morphophonological grammar* is necessary as an interface between the mental lexicon and the articulatory and perceptual systems. In simple terms, we might refer to these as *representation* and *rules*,[1] respectively.

Let us consider a specific example. The vowel appearing in the plural (PL) suffix in Turkish is sometimes the back vowel [ɑ] and sometimes the front vowel [e], depending on the backness of the vowel to its left, as in (1); data from Kabak (2011).

(1) [dɑ̲l-lɑ̲r] 'branch'-PL
[je̲r-le̲r] 'place'-PL

One way to capture the fact that the PL suffix alternates between two forms is to represent its vowel as an abstract segment /A/ whose realization as [ɑ] or [e] must be predicted from the environ-

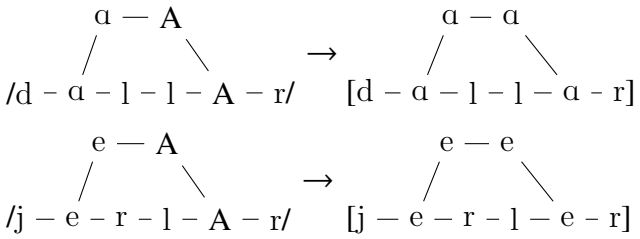
[1]We use the term *rule* here rather loosely to refer to any sort of generalization for mapping between underlying and surface forms, not as a commitment to rule-based formalisms over constraint-based formalisms. We might have said representations and *generalizations*, but that is rather a mouthful and misses an alliteration.

ment. This change of representation requires mapping between the abstract level of representation (denoted with '/-/') and the concrete level (denoted with '[-]'). This idea is indicated in (2).

(2) /dɑ̲l-lA̲r/ ⟷ [dɑ̲l-lɑ̲r]
/je̲r-lA̲r/ ⟷ [je̲r-le̲r]

In a flat, string data structure like (2), the interacting vowels are separated by intervening consonants. If the word is instead represented as a graph structure like (3) in which the vowels are projected onto a separate tier, then the process of /A/ warping into [ɑ] or [e] can be captured more naturally by interpreting the backness of the stem's vowel as spreading across the upper tier of the graph (Clements, 1976, 1980). Thus, the choice of representation has implications for the intrinsic character of the rules used to convert between levels of representation.

(3)



Phonological theory, at least that in the spirit of generative linguistics, has indeed occupied itself with the study of phonological rules and representations. The enormously influential work of Chomsky and Halle (1968) took a somewhat restrictive view of phonological representations, treating abstract representations as flat feature matrices and placing emphasis on "grammar as a system of rules" (Anderson, 2021, p. 464). Moreover, it was quickly made clear, in particular by Kiparsky (1973), that the null hypothesis ought to be that mental representations are concrete, and the positing of non-concrete mental representations should only be done when motivated. However, a convincing answer to the question of when abstract representations are justified has arguably never been established, and the debate sparked by Kiparsky was largely abandoned upon the arrival of new, representation-centric theories of phonology, which included autosegmental theory (Goldsmith, 1976), metrical theory (Liberman and Prince, 1977), and feature geometry (Clements, 1985).[2] McCarthy (1988, p. 84) described the spirit of these theories as follows: "if the representations are right, then the rules will follow."

The shift towards rich representations and the lack of a satisfactory justification for them has left an apparent degree of consternation. Some have argued for surface-driven theories (e.g., Albright 2002), and attempts to justify the positing of abstract representations in phonological theory have

[2]Our view of the history was influenced by the narratives in Goldsmith and Laks (2006) and Anderson (2021).

continued. Hyman (2018) argued that non-concrete underlying representations of words are justified for at least some phonological phenomena, provided they are not overly abstract, they allow for parsimonious characterization of the phenomena, they do not lead to unintended side effects, and they capture a speaker's productive knowledge of the phenomenon. Goldsmith and Riggle (2012) demonstrated that a model incorporating tiers provides better compression of linguistic data than one without, and argued that this provides statistical justification for their use in phonological theory. Others, have used the perspective of phonological learning to argue similarly. In particular, Hayes and Wilson (2008) presented empirical learning results that demonstrated the benefit of tier-based and grid-based representations for learning in the presence of dependencies that are non-local on a string representation, and Heinz et al. (2011); Jardine and McMullin (2017); Burness and McMullin (2019) provided learning-theoretic grounding to Hayes and Wilson's empirical observation by demonstrating how tier-based representations of words allow for strong theoretical guarantees on learning.

However, phonological learning has tended to use the metaphor of *search*. Under this metaphor, a phonological theory specifies a space of possible grammars. Each adult language is taken to be a particular grammar instantiated from that space, which produces primary linguistic data. This primary linguistic data forms the input to a learning algorithm, which in turn searches the space of grammars delineated by the phonological theory for the correct adult grammar. The most widely-adopted version of this stemmed from Optimality Theory (OT; Prince and Smolensky 1993). In an influential paper introducing learning in OT, Tesar and Smolensky (1998, p. 236) claim that "The general challenge of language acquisition, under any linguistic theory, is that of inferring the correct grammar from overt data." The space of grammars delineated by OT consists of all possible rankings of a universal constraint set—this is known as the *factorial typology*. Since constraints are taken to be universal and each language is one particular ranking of those constraints, learning in OT amounts to using primary linguistic data to figure out the correct constraint ranking. OT learning models are thus constraint-ranking or constraint-weighting models, such as Tesar and Smolensky (1998)'s Constraint Demotion algorithm, the Gradual Learning Algorithm (GLA; Boersma 1997; Boersma and Hayes 2001) for Stochastic OT, Boersma and Pater (2008)'s version of GLA for Harmonic Grammar, and Goldwater et al. (2003)'s maximum-entropy-based model (see Jarosz 2019 for an overview).

The metaphor of search and the conception of a "correct" grammar may have misled this approach, resulting in a misconstrual of the nature of learning. This metaphor presents a picture of learning in which the adult grammar is a terminal state and the child learner's objective is to discover the grammatical system underlying the ambient language to which they are exposed. This places the specification of a theory of phonological grammar as antecedent to a theory of learning, since this metaphor requires the structure of phonological grammars to be specified prior to

a learner commencing its search. Thus, the inclusion of a given representation in the theory of phonology cannot be a consequence of the learning algorithm. Even Hayes and Wilson (2008), who attempt to provide a learning-based justification for tiers, are constrained by the metaphor to positing two distinct theories of phonological representation to release their learning model upon: one with autosegmental tier representations and the other without. Only once these alternatives are posed are learning simulations able to suggest a choice between these alternatives.

The dissertation's epigraph suggests a different metaphor. We propose that language learning is better thought of as a continual process of construction. The process' terminal state is a fictitious idealization and the process' objective is not to reach it. The core intuition behind this dissertation is that a metaphor more appropriate than that of *search* is *construction*. This metaphor suggests that the learner is continually developing in their use of the language. The learner's starting point can be a fully-concrete lexicon in which words can be accessed directly, without a grammar,[3] but the Zipfian character of linguistic distributions will inevitably drive the learner to generalize from their linguistic experience.[4] Whereas work guided by the metaphor of search starts with the learner's input (training data), and posits the structure and content of its output (rules and representations) before working out the learning algorithm, work guided by the metaphor of construction starts with the input and hypothesizes the learning algorithm before working out what output (rules and representations) this suggests.

The algorithmic approach to phonology proposed in this dissertation thus starts by identifying independently-established psychological mechanisms that could be at play in this process of development. It then uses these mechanisms as the components of an explicit computational algorithm, which constitutes a hypothesis about the process of human learning. Through evaluation of the hypothesized learning algorithm—in particular its accuracy generalizing to unseen test words, its consistency with known developmental patterns, and its predictions in experimental settings—the algorithm can be interpreted as providing a learning-based account of the rules and representations that it outputs. These structures need no longer be posited *a priori* if they can be demonstrated to be the natural consequence of a learning algorithm that is grounded in independent psychological mechanisms.

In this dissertation, we demonstrate results enabled by this approach. In chapter § 2, we provide evidence that learners' starting point is indeed a concrete lexicon accessible by direct access, but that the sparsity of language necessitates extensive generalization. We also identify key psychological mechanisms that have independent motivation and serve as the building blocks of learning algorithms proposed in chapters § 3-5. Chapter § 3 proposes a learning-based account for the graph representations exemplified in (3) by showing that they are the natural consequence of an

[3]See § 2.3 for evidence.

[4]See § 2.2 for evidence.

algorithm grounded in the mechanisms identified in chapter § 2. In a similar fashion, chapter § 4 proposes a learning-based account for the abstract representations of morphemes, like the PL suffix in Turkish (2), by showing that effective generalization sometimes requires them. By the same token, the learning-based approach to underlying representations in chapter § 4 shows that effective generalization is sometimes possible without underlying abstraction, and this result allows for the novel interpretation of experimental results. Chapter § 5 zooms in on the process of constructing rules to map between abstract underlying representations and their concrete surface realizations, and demonstrates how our proposed learning-based approach can give greater clarity to the nature of typologically-rare processes. Chapter § 6 provides a novel experimental study directly testing the predictions of our proposed learning algorithms. We now turn to the dissertation's structure and a summary of its main contributions.

1.1 Contributions and Dissertation Structure

We have indicated that the core idea of this thesis is to directly study the processes by which language learners construct a phonological system. The starting point of this endeavor is to identify the psychological mechanisms available to this process. Chapter § 2 will provide relevant background for the rest of the dissertation, and, in particular, it will propose three psychological mechanisms that will form the basis of chapters § 3-6, whose main contributions we outline in Tab. 1.1 and summarize next (§ 1.1.1-1.1.3). The dissertation is highly interdisciplinary, with its contributions spanning both computer science and linguistics. In chapter § 7, we will review the main contributions—with the added clarity of coming at the end of the dissertation—along with a critical comparison of neural networks to our approach, and a discussion of limitations and future directions.

1.1.1 Change of Representation When and Only When Needed

Chapter § 3 presents a learning algorithm that attempts to predict the surface form of an alternating segment by tracking the segments adjacent to it. When this fails, the model deletes segments unhelpful to generalization, yielding a new representation. This step is repeated iteratively until the relevant dependencies for predicting the segment become local on the constructed representation. In this way, the model demonstrates how graph-like representations of words—usually called *autosegmental tiers* in phonological theory—can emerge as the computational consequence of a learning procedure grounded in humans' aforementioned proclivity for tracking adjacent dependencies. The model's iterative behavior precisely predicts human behavior in a range of artificial language experiments, and provides the model with sufficient flexibility to account for cross-linguistic variation. Moreover, because the model tracks adjacent dependencies first, local phonological processes fall

out as a special case where no change of representation is needed.

In computational terms, this model operates over an input representation of strings, and automatically constructs from this a graph representation, where dependencies that were non-local in the input string representation are local in the graph representation. The model outperforms a neural language model on small amounts of training data (see § 3.5 and § 4.3.4), demonstrating that change of representation can improve performance when long-distance dependencies are present in small amounts of data.

Ch. 3 Main Contributions

- LING: Algorithmic, learning-based account of phonological tiers
- LING: Handles cross-linguistic variation of tiers
- CS: Allows local learning over a graph structure
- CS: High accuracy on small data

In the closely-related chapter § 4, we turn to the question of when abstract underlying representations of words, differing from their concrete surface realization, need to be constructed. It is precisely the construction of such abstract representations that necessitates the learning of a mapping between abstract and concrete levels of representation. In this chapter, we argue that, by default, mental representations of morphemes correspond homomorphically to their physical realization, requiring nothing but a trivial mapping to convert between levels. However, in some cases, the fact that words are distributed sparsely in linguistic samples[5] necessitates the positing of abstract underlying structure to enable effective generalization to unseen word forms. The chapter § 4 model uses the Tolerance Principle (Yang, 2016), described in § 4.2.4, to determine when such abstract representations are needed. As was the case for the § 3 model, a change of representation from concrete to abstract is made when and only when needed for effective generalization. The abstract underlying forms and their corresponding surface realizations then become the training input to the chapter § 3 model.

We evaluate the model on two case studies, one in Turkish and one in Dutch. We find that the combination of the chapter § 3 and chapter § 4 models accounts for a range of morphophonological complexities in Turkish while achieving high accuracy generalizing to held-out test words. Moreover, the model provides a plausible learning-based account of why Dutch-learning children appear to lack productive knowledge of a well-studied voicing alternation, despite the fact that children exhibit clear productive knowledge of many alternations in other languages.

The chapter § 4 results reinforce chapter § 3's conclusion about the importance of automatic representation change in effective learning on small amounts of data.

[5]See § 2.2 for a summary of the facts.

Ch. 4 Main Contributions

- LING: Algorithmic, learning-based account of when abstract underlying forms are needed
- LING: Predicts phonological productivity in some cases but not others
- CS: High accuracy on small data

1.1.2 Expansion of Attention When and Only When Needed

In Chapter § 5, we turn in more detail to the question of how phonological processes—mappings between abstract and concrete levels of representation—are constructed. We propose a learning algorithm that models attention as initially centered around a single, alternating segment. The model then incrementally increases the width of its attention, only when needed to account for the surface form of the alternating segment. As the model in § 3, this iterative expansion of attention follows the experimental evidence that learners only begin tracking non-adjacent dependencies when doing so is necessary. However, whereas the model in § 3 changes representation when adjacent dependencies are insufficient, the model in § 5 expands the width of its attention. This can be thought of as zooming in on the default case from § 3, where no change in representation is needed, or, more generally, as zooming in on the process of constructing a local[6] generalization over a representation, whether that be a string or a tier. In § 7.3.1 we discuss in more detail how these two models fit together.

When compared to a number of alternative models for learning phonological processes, our proposed model is the only one to match human behavior in artificial language experiments. In particular, our model mirrors the fact that humans more easily learn a generalization to predict the surface form of an alternating segment when it is determined by a segment two steps away, than when it is determined by a segment three steps away.

Moreover, our model is able to construct rules for phonological processes that appear to work in opposition to clear articulatory motivation, but sometimes arise in languages due to historical contingency (Beguš, 2018). Traditional theories of phonological learning assume substantive constraints against marked sequences, and predict that such processes should be unlearnable. Nevertheless, such processes exist, are productive (Coetzee and Pretorius, 2010), and are readily learned in experimental settings (Seidl and Buckley, 2005; Beguš, 2018).

This model shows promise for improving the sample-efficiency of neural language models. As the model in chapter § 3, the model in chapter § 5 outperforms a neural network model on small amounts of training data (see § 5.4). The insight of this model—that attention can be incrementally expanded only as needed—is applicable to language modeling with a convolutional neural network

[6]Here *local* means that the surface form of an underlying segment is predicted from segments within a fixed distance of it on the representation. See § 5.2.2.1 and § B.1.2 for more details.

architecture, because convolutional layers involve a kernel size that determines the range of the model's attention. Thus, incrementally increasing the kernel size during training could lead to increased sample-efficiency, as we discuss in § 7.2.6.

Ch. 5 Main Contributions

- LING: Learning-based account of local phonological rules
- LING: Algorithmic explanation for human bias towards more local generalizations
- LING: Account of learning alternations lacking phonetic motivation
- CS: Incremental change of attention
- CS: High accuracy on small data

1.1.3 Experimental Validation of Model Predictions

Chapter § 6 evaluates the principle at the core of the computational models presented in § 3 and § 5, namely that learners do not track non-adjacent dependencies if an alternation can be predicted from adjacent dependencies.

The results demonstrate that attention is not globally applied and then narrowed to the relevant segments, as in a transformer model. Rather, attention starts locally and, when this is sufficient for generalization, learners show no evidence of sensitivity to other, equally-robust statistical dependencies. This contributes novel experimental evidence supporting our proposed idea that change of representation and broadening of attention is *incremental*, which has implications for understanding the order in which learners construct generalizations and for sample-efficient learning on small amounts of data.

This dissertation is built on the idea that an appropriate and productive way to study the nature of phonological representations and generalizations is to directly investigate the computational procedures by which a phonological system is constructed in the mind. This places computational learning procedures as the central object of study, and works from these up to the phonological structures that they construct. As a result, this dissertation complements work on a formal-language-theoretic approach to phonological representations and generalizations (see Heinz 2018 for a survey). That line of work places phonological structures as the primary object of study, by evaluating their computational properties and categorizing them in terms of these, and then interprets what constraints this puts on any algorithms that might construct these structures.

The way in which our approach complements formal language theory becomes especially clear when considering the results to the chapter § 6 experiment. The results provide novel insight into the order in which generalizations are constructed; insight that formal-language-theory is not well-suited to provide in this case because multiple generalizations consistent with the training data have

Table 1.1: Summary of Contributions

Contribution	Chapter(s)	Field(s)
Importance of Learning *Algorithms*	3,4,5,6	CS + LING
Insight into Human Generalization Behavior	3,4,5,6	LING
Efficient Learning with Small Data	3,4,5	CS + LING
Cross-Linguistic Variation	3,4,5	CS + LING
Change of Representation and Abstraction	3,4	CS + LING
Expansion of Attention	5	CS + LING
Non-Locality is Systematic	3	CS + LING
Unification of Locality and Non-Locality	3	CS + LING
Understanding of the Typologically Rare	5,6	LING
Insight into Dutch voicing alternations	4	LING
Account of Turkish Complexities	3,4	LING

the computational structure attributed to natural-language phonological processes (§ 6.5.1).

The experiment also contributes to our understanding of the nature of linguistic generalization, as it provides a case in which the most *local* generalization is not the most *conservative* generalization. Conceptual, learning-theoretic arguments have often led linguists to posit that learners generalize as conservatively as possible (Berwick, 1985; Albright and Hayes, 2003; Hale and Reiss, 2008), but our experiment results suggest that this is not always the generalization strategy.

Ch. 6 Main Contributions

- LING: Experimental validation of chapters § 3 and § 5
- LING: When generalization is not maximally conservative
- CS: Confirmation that movement away from locality is incremental

CHAPTER 2

Background

We begin by describing morphophonological alternations and the topics surrounding them that will feature prevalently in the dissertation (§ 2.1). We then turn to the sparsity of linguistic distributions, and how this necessitates generalization (§ 2.2). We will then present evidence that the learner's starting point is one in which words are stored concretely, allowing for direct access to lexical entries (§ 2.3). Next, we present experimental studies on sequence learning, which overwhelmingly reveal that a learner's attention is by default concentrated on elements adjacent in a representation (§ 2.4). Lastly, in § 2.5 we introduce the Tolerance Principle, which provides a precise, cognitively-grounded calculus for when the problems of sparsity, reviewed in § 2.2, no longer allow learners to operate under the defaults in § 2.3-2.4.

2.1 Morphophonological Alternations

Phonological segments often alternate in a way that is predictable from the phonological environment. In many cases, the relevant dependency conditioning the alternation is between adjacent segments. This is the case, for example, with the English plural suffix, which alternates between [-z] and [-s], matching the voicing of the stem-final segment—to which it is adjacent—and separated by [ə] if the stem-final segment is a sibilant. This is shown in (4).

(4) [dɑgz]
[kæts]
[hɔrsəz]

In some cases, it is a stem that alternates. For instance, Dutch exhibits a restriction against obstruents in syllable-final position. This results in an alternation in some singular-plural paradigms where a stem-final obstruent is voiced in the PL form but voiceless in the SG form, because there it is in final position (5a). In other paradigms, stem-final obstruents are voiceless throughout (5b).

(5) a. [bɛt] 'bed' [bɛdən] 'bed-PL'

b. [pɛt] ‘cap’ [pɛtən] ‘cap-PL’

These alternations have phonetic motivation. For example, in (4), assimilating the voicing of the suffix to the final segment of the stem avoids the need for a change in the glottis, and the separation of sibilants with a vowel avoids overly-similar adjacent sounds. Not all alternations enjoy such phonetic motivation. Tswana exhibits post-nasal devoicing of labial stops (Cole, 1955; Schaefer, 1982; Coetzee and Pretorius, 2010), even though post-nasal articulation promotes voicing of stops, and the opposite process—post-nasal *voicing*—is typologically pervasive (Locke, 1983; Rosenthall, 1989; Pater, 1999; Hayes and Stivers, 2000; Beguš, 2016, 2019). The resulting alternation can be seen when the 1ST SG OBJ clitic, which is a nasal, leads to devoicing of a stem-initial stop (6a) that is voiced in other contexts, such as when joined with the non-nasal /re-/ 1ST PL OBJ clitic (6b); data from Coetzee and Pretorius (2010, p. 406).

(6) a. [m-patla] ‘want me’
[m-potsa] ‘ask me’
[m-pulela] ‘open (for) me’

b. [re-batla] ‘want us’
[re-botsa] ‘ask us’
[re-bulela] ‘open (for) us’

Alternations appear in linguistic data as a distributional regularity. It does not automatically follow from this that a speaker necessarily has productive knowledge of the alternations distributionally present in their language. We turn to this issue next.

2.1.1 Productivity, Underlying Forms, and History

Productivity refers to a speaker’s ability to extend a distributional pattern to novel linguistic structures. The classic test of productivity in morphophonology was introduced by Berko (1958), and is called the *Wug Test*. Berko’s study, which investigated English morphophonology, involved presenting young children with a drawing of a made up animal—most famously a *wug* [wʌg]—along with the linguistic information that “This is a wug.” The child was then shown a drawing of two of the same animal and told “Now there is another one, there are two of them,” before being asked “there are two __?” If the child responded to the query with [wʌgz], this demonstrated that they have productive knowledge of the morphophonology of the English plural: they can extend it to nonce words. Berko (1958)’s paradigm is now the standard framework used in experimental studies of productivity, though the details differ depending on the phenomena being evaluated.

Berko (1958)’s study demonstrated that the English plural alternation exemplified in (4) is productive for young learners of English. Despite Tswana’s post-nasal devoicing (6) operating against

apparent phonetic motivation, Coetzee and Pretorius (2010) employed a Wug Test and found the process to be productive for some speakers. However, not all alternations that have been investigated show evidence of productivity. Zamuner et al. (2006, 2012) performed both production-based and comprehension-based Wug Tests on Dutch-learning children to test whether they have productive knowledge of the voicing alternation described in (5). They found no evidence of productivity. Thus, productivity is observed for some alternations, but not for others.

A common view is that language learners initially store words concretely, as accurately as their representational capacities allow (e.g., Hale and Reiss 2008; Ringe and Eska 2013; Richter 2021). In § 2.3, we summarize experimental, historical, and lexical studies supporting this view. As the child's lexicon grows, surface alternation—like the multiplicity of plural morpheme forms [z], [s], and [əz] in (4)—violates mutual exclusivity, which is the principle of one morpheme form per meaning. Since the surface form of the morpheme is predictable from the phonological environment, the violation of mutual exclusivity can be side-stepped by creating an underlyingly abstract plural morpheme together with productive processes for producing the appropriate surface form of the plural morpheme when it attaches to a stem, as illustrated in (7) where /-z/ denotes the abstract plural morpheme.[1]

(7) /dɑg-z/ → [dɑgz]
/kæt-z/ → [kæts]
/hɔrs-z/ → [hɔrsəz]

Under this view, it is when, and only when, concrete segments are collapsed into abstract underlying representations that the need for a phonological grammar arises, to derive the surface form for abstract underlying forms. Thus, lack of productivity indicates that the amount of surface alternation may not be pervasive enough to drive the learner to construct an abstract underlying form and a productive process to predict its surface realizations. In chapter § 4, we begin developing a learning-based answer to the question of when non-concrete underlying forms are constructed, which begins to bring some clarity to why the English plural alternation (4) appears to be productive while the Dutch plural alternation (5) appears not to be. In chapters § 3 and § 5, we propose learning algorithms that construct productive processes to map these underlying forms to their surface realizations.

The view of productivity as driven by prevalence of surface-alternation also has implications for the relationship between phonological learning and *diachrony*, or language history and change. Recall that the post-nasal devoicing alternation in Tswana operates in apparent opposition to phonetic motivation, yet Coetzee and Pretorius (2010) found evidence that it is productive for some speakers.

[1]It is conventional to treat this underlying form as we do here, with /-z/ devoiced into [-s] or separated by an epenthetic [ə]. However, we could also treat it as the even more abstract /-Z/, whose surface realization as [-z], [-s], or [-əz] is underdetermined by the lexical representation.

This is a puzzling result if productive phonological processes are expected to have a phonetic basis, but less so if productivity is the result of surface alternation. If phonological processes, productive or not, do not necessarily have a direct, phonetic origin, then whence the process?

Historical linguistics provides a likely answer. Beguš (2019, p. 699) found post-nasal devoicing to be reported as a process in thirteen languages and dialects, and argued that despite being unnatural, it likely emerged in each case as the result of a sequence of sound-changes. Each of the sound-changes was phonetically motivated given the previous one, but their composition—post-nasal devoicing—cannot itself be easily provided a phonetic justification. This sort of historical process, sometimes called *telescoping*,[2] has occurred reasonably often in the world's languages (Kenstowicz and Kisseberth 1977, ch. 2; Johnsen 2012). Experimental evidence supports this view, as participants readily learn phonological processes lacking phonetic motivation in studies by Seidl and Buckley (2005) and Beguš (2018), as well as our own experiment in chapter § 6.

Our treatment of productivity as being driven by prevalent surface-alternation thus has the potential to give an explanation for how it is that some Tswana learners were able to construct a productive process of post-nasal devoicing, and why a similar result is observed in experimental settings. We return to this topic in chapter § 5.

2.1.2 Non Adjacency

Once a non-concrete underlying form is constructed, learners must construct a phonological process to predict the various surface realizations of the abstract underlying form. While the alternations described in (4)-(6) involved interactions between two adjacent segments, this is far from a universal property. In some cases, it is necessary to track dependencies at a greater distance. For example, in some dialects of English, a segment that usually surfaces as [t] or [d] is instead realized as the alveolar tap [ɾ] in intervocalic position, if the following vowel is unstressed. This leads to alternations like (8), where a [ɾ] replaces [t] when appearing between vowels and the second vowel is unstressed.

(8) [it] 'eat' [iɾɪŋ] 'eating'

If the underlying segment is an abstract segment /T/, as appears to be the case (Richter, 2018, 2021), then predicting the surface realization requires tracking dependencies on both sides of the underlying segment, not just a single adjacent segment. In chapter § 5 we propose a computational model that provides a hypothesis about how productive processes are constructed to account for alternations involving dependencies within some fixed distance of the alternating segment.

[2]Presumably this term is meant to elicit the mental imagery of a telescope sliding into itself, where each sound change in the chronological sequence forms one unit of the telescope. The metaphor has the added benefit of capturing the problem-solution relationship between the fact that the resulting, unmotivated process obscures its historical origins, and that telescopes are useful for bringing into view the difficult-to-see.

More dramatically, some alternations—particularly those occurring in consonant harmony (Rose and Walker, 2004), consonant dissimilation (Bennett, 2013), and vowel harmony (Van der Hulst, 2016)—involve dependencies between segments that are arbitrarily far away. For instance, the vowels of Turkish suffixes harmonize with the preceding vowel across varying numbers of intervening consonants. In (9), the affix vowels alternate between back {ɑ, ɯ} and front {e, i} to match the [back] value of the preceding vowel.

(9)	[dɑ̱l-lɑ̱r-ɯ̱n]	branch-PL-GEN	(Kabak, 2011, p. 3)
	[j̱e̱r-le̱r-i̱n]	place-PL-GEN	(Kabak, 2011, p. 3)
	[i̱p-le̱r-i̱n]	rope-PL-GEN	(Nevins, 2010, p. 28)

We mentioned in the introduction that Turkish vowel harmony could be viewed as a process of spreading across a vowel tier in a graph representation of a word. This has been a common analysis since the development of autosegmental tiers (Goldsmith, 1976; Clements, 1976, 1980). However, in chapter § 3, we demonstrate that the graph needed to view processes in this way varies cross-linguistically. In the same chapter, we then provide a computational model as an explicit hypothesis about how the mind may construct these non-flat representations of words.

2.2 Sparsity in Language

2.2.1 The Frequency Distribution of Language Requires Generalization

The statistical distribution of words in a corpus of reasonable size consistently follows Zipf (1949)'s "law," which states an inverse relationship between the rank of a word type and its frequency. Specifically, Zipf's law predicts that the second most frequent word (rank 2) is roughly one half as frequent as the most frequent word, the third most frequent word (rank 3) is roughly one third as frequent as the most frequent word, etc. This relationship can be seen by plotting, in log scale, the frequency of a corpus' word types as a function of their rank frequency. For example, such a plot can be seen in Fig. 2.1, which we derived from the child-directed speech in Roger Brown (1973)'s acquisition study of English.

The substantive implication of this robust empirical observation is that most of a language's words occur very infrequently and only a few occur frequently. It follows from this, that children's early mental lexicons—over which they learn the core of their language's morphophonology—contain only a tiny fraction of the entire vocabulary of the language. Thus, in order to effectively use the long-tail of language, children must aggressively generalize from their early mental lexicons.

2.2.2 The Indefinite Size of Language Requires Generalization

The other major statistical observation about corpuses is Heap's law (sometimes called Herdan's law) (Herdan, 1960), which gives the number of word types as a function of number of tokens, and has the intuitive interpretion of *diminishing returns*: As the number of tokens increases, the number of types increases, but at a decreasing rate. However, this growth function does not have an asymptote. It thus appears that vocabulary size does not have a non-arbitrary limit (Chan, 2008, p. 24). This can be seen in Fig. 2.2, where the number of word types in Brown (1973)'s child-directed corpus increases rapidly for the first tokens, but the rate of new types slows down as the number of tokens increases, while never appearing to approach an asymptote.

This slowing down of new word types entering the child's mental lexicon after the most frequent types corroborates the implications of Zipf's law: children must generalize from their early mental lexicons in order to accommodate the new words that gradually and persistently enter at later points.

2.2.3 Paradigm (Un)saturation Requires Generalization

A major manifestation of the fact that most word types occur very infrequently and new words enter the vocabulary at decreasing rates is that morphological paradigms are generally incomplete. Chan (2008) introduced the concept of *paradigm saturation*, which measures, for each lemma, the fraction of the total number of inflectional categories in the corpus for which the lemma's inflectional form is attested in the corpus (Yang, 2016, p. 21). For example, English verbs have six inflectional categories, as exemplified with the verb *walk* in (10).

(10) Infinitive: To <walk>

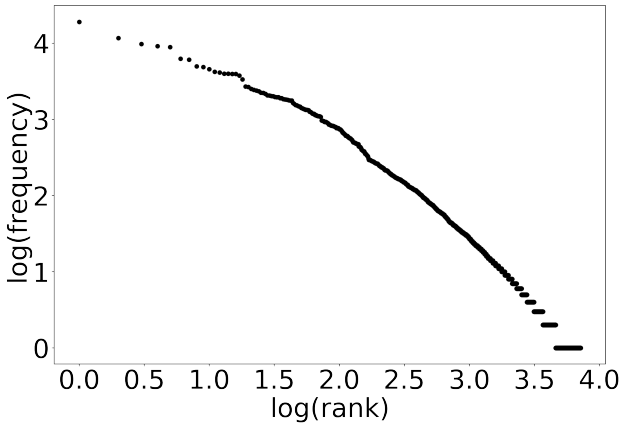


Figure 2.1: The log frequency of child-directed word types in Brown (1973)'s acquisition corpus approximately follows the famous Zipfian distribution, in which a few words occur very often and most words occur only a few times.

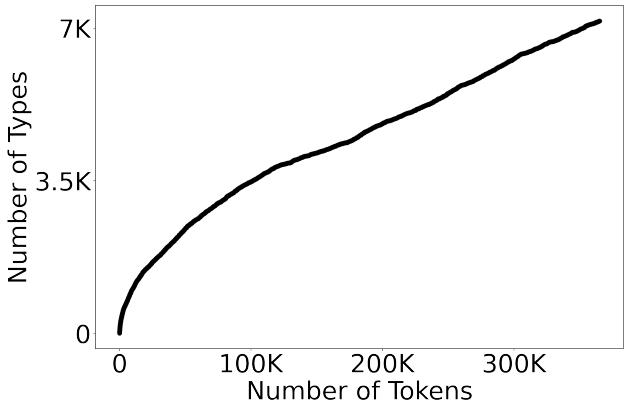


Figure 2.2: The rate of new child-directed word types in Brown (1973)'s acquisition corpus decreases as the number of tokens increases, but never approaches an asymptote.

First and second person present tense: I/you <walk>
Third person singular present tense: She <walks>
Progressive: They are <walking>
Past tense: They <walked>
Past participle: They have <walked>

If a corpus (or a child's mental lexicon) contains, e.g., four of these six forms, then the paradigm saturation for the lemma <walk> would be 2/3.

Chan (2008, Tab. 4.4) reports that the average saturation across 16 corpuses of different languages was only 71.3%. The situation seems quite dire for languages with rich morphologies. For example, Chan's table reports the *maximum* saturation of any stem in Finnish—an agglutinative language—to be 40.3%. This makes concrete the abstract observation from § 2.2.1-§ 2.2.2 that the sparsity of language requires the child to generalize from their early mental lexicon to the long tail of the linguistic distribution. Suppose a child knows some singular and some plural nouns, as shown in Tab. 2.1. For some words, the child knows both the singular and the plural forms (e.g., *dog* and *dogs*); for others they know only one (*cat* but not *cats*). At this stage, a word like [kæt] has the same intrinsic status as *Wug* in Berko (1958)'s seminal study: namely, the child knows the SG, but to use the PL must produce a novel form. Here, the motivation for predicting the plural form arises on the occasion of needing to use that form but finding an empty paradigm cell. The statistical profile of language—in particular the results from Chan (2008)—demonstrate that such a scenario is pervasive, and perhaps the driving force behind the construction of a morphophonological grammar.

Table 2.1: A hypothetical state of a child's mental lexicon for some English nouns. The child only knows both the SG and PL forms for some words.

SG	PL	Gloss
[kæt]	??	cat ~ cats
[dɑg]	[dɑg-z]	dog ~ dogs
[hɔrs]	[hɔrs-əz]	horse ~ horses
??	[bɝd-z]	bird ~ birds
[wɛb]	??	web ~ webs
[seɪf]	[seɪf-s]	safe ~ safes
??	[mæp-s]	map ~ maps

2.2.4 Summary

The previous subsections demonstrate that the empirical conditions of language acquisition require children to generalize beyond their early vocabulary, thus constructing a productive grammar. In § 2.5 we present Yang (2016)'s Tolerance Principle, which provides a precise, cognitively-grounded tipping point at which rote memorization becomes less cognitively efficient than generalization. The derivation of the tipping point, presented in Yang (2016, ch. 3) is grounded in the statistical profile of language data reviewed above.

2.3 Studies Pertaining to Lexical Representations

The view that representations of words are, at least initially, minimally abstract receives support from a range of theoretical perspectives. Kiparsky (1968) observed that in the absence of alternation, children have no reason for constructing an abstract representation of a morpheme. This has been incorporated into Optimality-Theoretic-approaches to learning, which usually posit that children's starting point is to construct underlying forms identical to their observed surface realization (Hayes, 2004; Tesar, 2013). The same position is proposed by Ringe and Eska (2013), who observe that historical sound changes support the position. Below, we interpret lexical and experimental studies, which we view as providing support for this widely-adopted position.

2.3.1 Lexical Studies

In a detailed study of the Providence corpus (Demuth et al., 2006) of the CHILDES database, Richter (2018, 2021) studied the acquisition of the English flap [ɾ] as an allophone of an abstract underlying /T/ phoneme. Richter found a U-shaped development curve, which is characteristic of the construction of a linguistic generalization. Up to 3;0, children produce the flap [ɾ] in the

intervocalic contexts that it would be expected in (i.e. where the following vowel is unstressed). However, between 3;0 and 5;0, children show in increased rate of producing *[t] instead of [ɾ] in those contexts. The children eventually return to accurate production of [ɾ] in the expected contexts by around the time they learn to read.

It thus appears that children are initially positing underlying /ɾ/ for words with surface [ɾ]. The fact that children begin to produce *[t] in lieu of [ɾ] suggests that children eventually construct an abstract underlying form that either includes both [t] and [ɾ] (e.g., /T/), or that replaces /ɾ/ with /t/, such that contexts with surface [ɾ] must be derived by a productive process.

2.3.2 Experimental Studies

Studies of early infant linguistic development have found that infants show an initial sensitivity to contrasts (e.g. [b] ~ [p]) regardless of whether a given contrast is linguistically significant in their native language, but that the sensitivity to non-native contrasts declines with age and linguistic experience (Werker and Tees, 1984; Kuhl et al., 1992). This change appears to be mediated by linguistic experience, which improves the recognition of acoustically non-salient native contrasts (Narayan et al., 2010) and does not degrade perception of non-native contrasts that do not conflict with the child's developing phonology (Best et al., 1988).

The ability to detect two sounds as distinct requires that children record the acoustic feature(s) that differentiate them. This supports the conclusion that children's phonetic representations initially record acoustic information concretely. The fact that the shift in perception of contrasts with age is mediated by linguistic experience suggests that organization into abstract structures is not the default, but rather a response to linguistic exposure.

When we turn to experiments pertaining to knowledge of alternations, results suggest that learners' default is to assume that novel words do not alternate. Dutch, in particular, has received a great deal of attention. As discussed in § 2.1, some Dutch noun paradigms exhibit a voicing alternation in which a stem-final obstruent is voiced in the PL form but voiceless in the SG form, where it occurs in final position (e.g., [bɛt] 'bed' ~ [bɛdən] 'bed-PL').

Suppose a Dutch-learning child is presented a nonce plural [slɑd-ən]. Under the hypothesis that children store words in their mental lexicon concretely by default, the child will construct /slɑd/ as the underlying form of the root. If the child is then asked to "Point to the [slɑt]" or to fill in the blank "There is one __," they may have difficulty doing so, because their concrete underlying form /slɑd/ does not match [slɑt] and violates Dutch's phonotactic restriction against voiced obstruents in final position. In particular, they should have greater difficulty completing these requests than if they are presented a nonce plural like [kɛt-ən], which does not alternate.

Zamuner et al. (2006, 2012)'s experiments tested precisely these predictions. We describe these

experiments in greater detail in § 4.4, where our model for learning underlying forms provides a novel interpretation for their results. To summarize the relevant results, Zamuner et al. found that Dutch-earning children at ages 2;6-3;6 indeed had greater difficulty comprehending and producing the singular form of nonce nouns when the plural form ended in a voiced obstruent than when it ended in a voiceless obstruent. As described above, this result is consistent with the predictions of the hypothesis that children initially store words concretely, since the plural forms with a voiced obstruent entail an alternation between the singular and plural that those with a voiceless obstruent do not—an alternation that cannot be captured with concrete underlying forms.

Coetzee (2009) performed a similar experiment in an artificial language experiment, and its results corroborate the natural language results from Zamuner et al. (2006, 2012). Coetzee's artificial language alternation was inspired by another alternation Dutch. In some noun paradigms, a vowel that is short in the SG lengthens in the plural, when a stem-final consonant is re-syllabified as the onset of the syllable containing the plural suffix, which leaves the vowel in a stressed, open syllable. Examples from Coetzee, p. 110 are shown in (11a). However, not all monosyllabic SG nouns with short vowels exhibit the alternation, as shown in (11b), and whether a noun is alternating or non-alternating is not phonologically-predictable.

(11) a. [xɑt] 'hole' [ˈxaː.tən] 'holes'
[spɛl] 'game' [ˈspeː.lən] 'games'
b. [kɑt] 'cat' [ˈkɑ.tən] 'cats'
[stɛl] 'set' [ˈstɛ.lən] 'sets'

When presented with singular-plural pairs from an artificial language modeled after the Dutch vowel-lengthening alternation, in which half of the words alternated and half did not, Coetzee found that learners were able to learn which words alternated, but did not extend the alternation to novel words. However, when a substantial majority of the exposure data were alternating paradigms, the learners did begin to extend the alternation to nonce paradigms, suggesting that sufficient evidence of surface alternation can lead learners to abandon the default creation of concrete underlying forms.

Together, these natural language and artificial language experimental results suggest that learners by default construct concrete underlying forms, but appear to move away from this default when substantial amounts of surface alternation are present in the linguistic data they are exposed to.

2.4 Experimental Studies of Sequence Learning

Early studies of statistical sequence learning were focused on the question of whether infant learners could segment a continuous speech stream into discrete units based on statistical information alone (Saffran et al., 1996, 1997; Aslin et al., 1998). Based on the idea that sounds crossing word

boundaries may co-occur less often than sounds within a word, these studies hypothesized that infants could track *transitional probabilities* and use them as a cue to segment speech. In their simplest form, transitional probabilities, which gained attention through Shannon (1948)'s Information Theory and have a long history is psycholinguistics (Levelt, 2013, ch. 12), capture the strength of dependency between adjacent segments in a set of sequences. The definition of the transitional probability from X to Y, denoted $\Pr(Y|X)$ is provided in (12).[3]

(12)

$$\Pr(Y|X) \triangleq \frac{\text{Freq}(XY)}{\text{Freq}(X)} \tag{2.1}$$

The numerator will be large if the sequence XY is frequent, and the denominator will be large if the sequence X is frequent. Since the sequence XY cannot occur without X occurring, and, in the limit, XY could never occur, the transitional probability is bounded by $0 \leq \Pr(Y|X) \leq 1$, as one would expect for a well-defined probability. Its maximum is achieved when every occurrence of X is followed by a Y—i.e. Y is totally dependent on X.

Consider the example set of sequences in (13a) and the calculations of $\Pr(Y|X)$ and $\Pr(Y|A)$ in (13b) over this data.

(13) a. XYZ
AYZ
ZXY
AXZ

b. $\Pr(Y|X) = \frac{\text{Freq}(XY)}{\text{Freq}(X)} = \frac{2}{3}$

$\Pr(Y|A) = \frac{\text{Freq}(AY)}{\text{Freq}(A)} = \frac{1}{2}$

Since $2/3 > 1/2$, this indicates that the strength of dependence between Y and X is greater than that between Y and A. To reiterate the main point: transitional dependencies of this sort capture the strength of dependence among segments, and, as defined in (12), the strength of dependence between *adjacent* items.

Statistical learning studies were also directed towards non-adjacent dependencies (Santelmann and Jusczyk, 1998; Newport and Aslin, 2004; Gómez, 2002; Gómez and Maye, 2005), because such dependencies seem relevant to learning syntax and morphology. As we discuss in § 2.1, non-adjacent dependencies are also pervasive in phonology. These studies usually investigated the ability to determine grammaticality of strings in an artificial language in which, for example, X is allowed to follow A across an intervening Z, and Y can follow B across an intervening Z (14a), but Y is not allowed to follow A, nor can X follow B (14b).[4]

[3]Higher-order transitional probabilities can be defined analogously.

[4]As is standard in linguistics, the symbol '*' denotes *ungrammatical* structures.

(14) a. $\underline{A}Z\underline{X}$
$\underline{B}Z\underline{Y}$
b. *$\underline{A}Z\underline{Y}$
*$\underline{B}Z\underline{X}$

In the remainder of this section, we will argue that the ability to track adjacent dependencies is an independently-established psychological mechanism that learners bring to the table. The argument rests upon the domain-independent observation of sensitivity to adjacent dependencies (§ 2.4.1), the asymmetrical developmental trajectory of adjacent vs. non-adjacent dependencies (§ 2.4.2), and learner's reluctance to track non-adjacent dependencies, doing so only as a final resort (§ 2.4.3).

2.4.1 Domain-Independence of Adjacent Dependency Tracking

The early studies of statistical learning applied to segmenting continuous speech streams found that infants as young as 8-months old were sensitive to dependencies between adjacent elements (Saffran et al., 1996, 1997; Aslin et al., 1998). The evidence came from the infants' ability to rapidly segment continuous speech streams that were carefully designed such that the relative size of transitional probabilities served as the only apparent cue to word boundaries.

The relevant transitional probabilities in these studies needed to be computed over syllables. However, the ability to track adjacent dependencies has also been attested when the transitional probabilities involved morphemes (Santelmann and Jusczyk, 1998), non-linguistic tones (Saffran et al., 1999), and visual shapes (Fiser and Aslin, 2002). The diversity of types of elements for which this ability has been observed suggests that the ability to track statistical dependencies between adjacent segments is neither limited to a particular kind of phonological structure nor even the domain of language.

2.4.2 The Developmental Trajectory of Sensitivity to Adjacent and Non-Adjacent Dependencies

While the ability to track adjacent dependencies has been widely attested, even for infants as young as 8 months old, Santelmann and Jusczyk (1998) found that even at 15-months-old, children showed no evidence of tracking dependencies between non-adjacent elements.

Studies with older participants revealed that the ability to track non-adjacent dependencies does eventually emerge: adults show a sensitivity to dependencies between non-adjacent phonological segments (Newport and Aslin, 2004), and 18-month-old children show sensitivity to dependencies between both non-adjacent morphemes (Santelmann and Jusczyk, 1998) and words (Gómez, 2002).

Figure 2.3: The developmental trajectory of learners' ability to track adjacent and non-adjacent dependencies. Infants as young as 8mo old show evidence of being able to track adjacent dependencies; this ability persists into adulthood. In contrast, do not show evidence of tracking non-adjacent dependencies until around 15mo.

Gómez and Maye (2005) attempted to map the developmental trajectory of this ability to track non-adjacent dependencies, and found that it grew gradually with age. At 12 months, infants showed no evidence of tracking non-adjacent dependencies, but they began to do so by 15 months, and showed further advancement at 17 months.

The developmental trajectory of adjacent and non-adjacent dependencies is summarized in Fig. 2.3. Infants as young as 8-months-old show evidence of tracking adjacent dependencies, and this ability persists into adulthood. Asymmetrically, infants only begin to show evidence of tracking non-adjacent dependencies at around 15 months.

2.4.3 Adjacency as Default

Even as sensitivity to non-adjacent dependencies develops, learners still more readily track local dependencies. Gómez (2002) found that 18-month-olds could track non-adjacent dependencies, but that they only did so when adjacent dependencies were unavailable. Gómez and Maye (2005) replicated these results with 17-month-olds and described the situation like this (p. 199): 'It is as if learners are attracted by adjacent probabilities long past the point that such structure is useful.'

Indeed, artificial language experiments have repeatedly demonstrated that learners more easily learn local phonological processes than non-local ones (Baer-Henney and van de Vijver, 2012) and, when multiple possible phonological generalizations are consistent with exposure data, learners systematically construct the most local generalization (Finley, 2011; White et al., 2018; McMullin and Hansson, 2019). In short, these studies demonstrate that learners posit the most local generalization consistent with the data.

2.4.4 Summary: A Proclivity for Adjacency

Zoologists usually recognize that different animals have different nervous systems, and that these differences shape—to some degree or another—each animal's behavioral and cognitive abilities, including their learning abilities (Gallistel, 2000). For acoustic processing, Podlipniak (2017, p. 2) puts it like this: 'every species is sensitive to specific acoustic cues due to the proclivities of its own nervous system.'

The evidence that infants can track adjacent dependencies across a range of items not limited to the linguistic domain (§ 2.4.1), that the ability to track non-adjacent dependencies appears to emerge later in development than that for adjacent dependencies (§ 2.4.2), and that learners only show evidence of tracking non-adjacent dependencies as a last resort when adjacent dependencies fail them (§ 2.4.3), strongly supports the conclusion that human's exhibit a proclivity for adjacency. This proclivity constitutes an independent foundation for the models presented in chapters § 3 and § 5.

2.5 The Tolerance Principle

The Tolerance Principle (TP), proposed by Yang (2016), is a cognitively-grounded tipping point, which hypothesizes that children form productive generalizations when the number of exceptions to a proposed generalization results in a real-time processing cost lower than that without the generalization. The tipping point depends on two values: the generalization's scope—that is, how many items it applies to—and how many exceptions the generalization has. It is based in the observation that the real-time computational cost of using a rule increases as the number of exceptions to the rule increases. The exact derivation of the TP is provided in Yang (2016, ch. 3), but rests critically upon the empirical observations of linguistic sparsity reviewed in § 2.2. Beyond its cognitive motivation, the TP has also had much prior success in computational modeling, lexical, and experimental studies (Schuler et al., 2016; Yang, 2016; Richter, 2018; Koulaguina and Shi, 2019a; Emond and Shi, 2021; Richter, 2021; Belth et al., 2021; Payne, 2022).

The threshold is stated in (15), where n is the size of the rule's scope, and e is the number of exceptions to the rule.

(15)
$$e \leq \frac{n}{\ln n}$$

When a generalization meets the threshold, the learner accepts it, and the exceptions to it can be lexicalized. An interesting property of the threshold $n/\ln n$ is that the relative number of exceptions tolerated decreases as n increases. Thus, as a learner's mental lexicon grows, in many cases n

will become larger and generalization will become more difficult. This intuitively matches the will-known fact that young children are more precocious language learners than older children and adults.

2.5.1 The Role of the Tolerance Principle

As discussed by Yang (2016, p. 10), and implemented by Belth et al. (2021), the TP can serve as an evaluation metric for the proposal and verification of rules in language acquisition. In the context of this dissertation, the TP provides an independently-motivated mechanism for abandoning the defaults discussed above in § 2.3 and § 2.4. In § 3, the TP will decide when adjacent segments unhelpful to generalization must be deleted. In § 4, the TP will determine when concrete underlying forms are no longer sustainable and abstract underlying forms must be created. In § 5, the TP will govern when the model must extend its attention beyond adjacent dependencies. Specifically, these default generalizations are abandoned when the number of exceptions to them grows too large, as measured by the Tolerance Principle.

CHAPTER 3

An Algorithmic Account of Phonological Tiers

The material in this chapter has been presented at the 2022 ACL workshop CMCL*, the 2022 MidPhon conference, the 2022 Linguistic Society of America Conference, and the 2023 Penn Linguistic Conference. A version of the chapter is under review at a major linguistics journal.*

Phonologists have at times been motivated to posit abstract, graph-like representations of words, sometimes called phonological tiers, in order to more naturally capture long-distance phenomena like vowel and consonant harmony. However, the data structure needed to render dependencies adjacent varies cross-linguistically, and the abstract nature of these representations in comparison to flat, string-like representations has led phonologists to seek justification for their use in phonological theory. In this chapter, we propose an algorithmic, learning-based approach. Our proposed model is grounded in humans' strong proclivity for tracking adjacent dependencies. We demonstrate that a graph-like representation can emerge as the computational consequence of a simple learning procedure that is restricted to only tracking adjacent dependencies. When trained on small amounts of natural language data, the model achieves high accuracy generalizing to held-out test words, while flexibly handling cross-linguistic complexities like neutral segments and blockers. The model also makes precise predictions about human generalization behavior, and these are consistently borne out in artificial language experiments.

3.1 Introduction

Phonological theory, and linguistics more broadly, often attempts to interpret apparently long-distance dependencies as being adjacent on some representation. Morphophonological alternations, such as (4)-(6) in chapter § 2 often involve dependencies between segments that are already adjacent. We repeat (4) below as (16). These require no new interpretation for the relevant segments to be adjacent.

(16) [dɑgz]
[kæts]
[hɔrsəz]

However, the alternations that arise in consonant harmony (Rose and Walker, 2004), consonant dissimilation (Bennett, 2013), and vowel harmony (Van der Hulst, 2016) often involve dependencies between segments that are arbitrarily far away in a string representation. We have demonstrated this with Turkish in (9), and we repeat the facts here. The vowels of Turkish suffixes harmonize with the preceding vowel across intervening consonants. Examples in (17) show the affix vowels alternating between back {ɑ, ɯ} and front {e, i} to match the [back] value of the preceding vowel. Arbitrary numbers of consonants intervene.

(17)	[dɑl-lɑr-ɯn]	branch-PL-GEN	(Kabak, 2011, p. 3)
	[jer-ler-in]	place-PL-GEN	(Kabak, 2011, p. 3)
	[ip-ler-in]	rope-PL-GEN	(Nevins, 2010, p. 28)

A similar phenomena is at play in the Omotic language Aari, where sibilant harmony is at play. This is exemplified in (18) from McMullin (2016, p. 21) (adapted from Hayward 1990). Underlying /s/ (18a) surfaces as [ʃ] when it is preceded by a [−ant] sibilant at any distance (18b).

(18) a. /baʔ-s-e/ → [baʔse] 'he brought'
b. /ʔuʃ-s-it/ → [ʔuʃʃit] 'I cooked'
/ʒaʔ-s-it/ → [ʒaʔʃit] 'I arrived'
/ʃed-er-s-it/ → [ʃederʃit] 'I was seen'

The development of representation-rich theories allowed these to be re-interpreted as being adjacent on some more abstract representation. In particular *autosegmental* or *phonological tiers* (Goldsmith, 1976; Clements, 1976, 1980) reduce alternations like these to local interactions on an alternation-relevant tier. We demonstrated this idea in the dissertation's introduction. The new representation is a graph structure, in which a new tier is projected so as to contain a subset of the string's segments. In Turkish (17), the relevant dependencies are local on a projected [+vowel] tier. In Aari (18), a [+sib] tier renders the dependencies local.

The change of representation needed to allow these processes to be viewed as a local process over a tier can vary extensively cross-linguistically, and is not always describable with a simple natural class like [+vowel] or [+sib]. For instance Finnish, like Turkish, exhibits [back] vowel harmony. However, the vowels {i, e} are neutral: they neither participate in nor block harmony. This is exemplified in (19) (Ringen and Heinämäki, 1999, p. 305), where the vowel in the essive (ESS) case suffix alternates between back [ɑ] and front [æ] depending on the final harmonizing vowel of the

stem (19a), but passing over both consonants and neutral vowels: the affix harmonizes with [+back] [o] not neutral [−back] [i] in (19b).

(19) a. [pøytæ-næ] table-ESS
[poutɑ-nɑ] fine weather-ESS
b. [koti-nɑ] home-ESS

In order for this alternation to be reduced to dependencies between adjacent segments, both the consonants and the neutral vowels must be excluded from the tier.

In Latin, default /l/ (20a) dissimilates to [r] when preceded by /l/ across varying distances (20b), but the dissimilation is blocked by an intervening /r/ (20c) or an intervening [−cor] consonant (20d) (examples from Cser 2010, organized following McMullin 2016).

(20) a. *nav-alis* 'naval'
b. *popul-aris* 'popular'
lun-aris 'lunar'
c. *flor-alis* 'floral'
d. *pluvi-alis* 'rainy'
leg-alis 'legal'

Consequently, in addition to [l] and [r], the relevant tier must preserve the [−cor] consonants to block the dissimilation (McMullin, 2016; Burness et al., 2021).

It is clear that no small number of tier representations can render dependencies local across the world's languages, as these examples and others (McMullin and Hansson, 2016; Burness et al., 2021) demonstrate. Moreover, because of the abstractness of these representations in comparison to flat, string representations, phonologists have sought justification for using these representations in phonological theory. Hayes and Wilson (2008) proposed an *inductive baseline* argument. They developed a model for learning phonotactics[1] and demonstrated that it cannot learn non-local generalizations unless it is provided the relevant tier projection *a priori*. Hayes and Wilson concluded that tiers might be necessary for non-local learning, and viewed this results as empirical, learning-based evidence in favor of the use of tiers in phonological theory. Formal-language-theoretic and computational-learning-theory results demonstrates that restricting a learner's hypothesis space to tier-based generalizations allows for proving strong, theoretical learning results (Heinz et al., 2011; McMullin, 2016; Burness and McMullin, 2019). These provide theoretical support for Hayes and

[1]The surface restrictions implied by phonological processes; e.g., the vowel harmony process in Turkish implies a restriction against non-harmonized vowels on the surface.

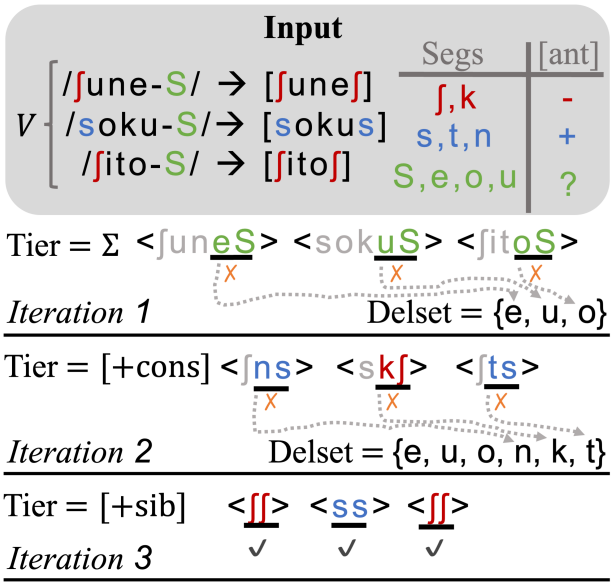


Figure 3.1: A visualization of our proposed algorithm running on a toy example of sibilant harmony.

Wilson (2008)'s empirical results. Goldsmith and Riggle (2012) demonstrated that tier representations allow for better compression of linguistic data, and viewed this as statistical justification for the use of tier representations.

In this chapter, we approach the problem from the opposite direction, using our algorithmic approach to demonstrate that tier representations are the natural consequence of a learning algorithm built from the results discussed in § 2.4, which demonstrated that humans have a proclivity for tracking adjacent dependencies.

We propose a computational model (§ 3.2) that tracks only adjacent dependencies, following the experimental evidence that humans track these more readily than non-adjacent dependencies (see § 2.4). When adjacent dependencies are not predictive of the alternation, the model is left to resort to deleting any adjacent segments that were unpreductive of the alternation. This suggests an iterative algorithm, in which the model iteratively tracks adjacent dependencies, deletes those that do not correctly predict the surface form of the alternating segment, and repeats on the new representation. This is visualized in Fig. 3.1, where we show the algorithm running on a toy example of sibilant harmony. On the first iteration, vowels are adjacent to the alternating sibilants. Because the vowels are underspecified for feature [ant], the sibilant cannot be predicted from them and they are deleted. In the second iteration, the consonants adjacent to the alternating sibilants predict the wrong surface forms, so they are deleted. Finally, in the third iteration, the sibilants are adjacent and can be accurately predicted.

In § 3.3, we discuss prior models for learning non-local phonological dependencies. We then compare our proposed model—and prior models—to human behavior on multiple prior artificial

language experiments in § 3.4. Our proposed model is the only one to match human behavior in all cases. In § 3.5, we evaluate our model at learning natural language, non-local alternations. Our model effectively learns Turkish vowel harmony, Finnish vowel harmony, and Latin liquid dissimilation, with substantially greater test accuracy than prior models. These results demonstrate that the model can handle parasitic harmony (Turkish secondary rounding harmony), neutral segments (Finnish vowel harmony), and blocking segments (Latin liquid dissimilation). The model achieves these results on datasets characteristic of the size of childrens' vocabularies at the time they appear to begin extending harmony to novel words (Altan, 2009). In chapter § 4, we use the model again on a second case study in Turkish, which corroborates the results in this chapter and demonstrates that alternations involving string-adjacent dependencies are a special case where no change of representation is needed. Together, these results suggest that the model provides a cognitively-plausible account of how the robustly-observed human proclivity for tracking adjacent dependencies can lead the mind to construct phonological tiers.

3.2 Model Description

Our model is called D2L for *Distant To Local* because it learns non-local alternations by automatically constructing a tier that exposes distant dependencies as local. D2L takes as input a set V of (UR, SR) input-output pairs. D2L computes the segments A that alternate in V, and then attempts to account for this alternation in terms of segments adjacent to the alternating segments. If this fails, D2L deletes the adjacent segments that failed to account for the alternation and tries again. This process iterates until an adequate generalization is discovered or until no more segments can be deleted.

In § 3.2.1, we describe the structure of D2L's generalizations: how they project a tier and how the surface form of alternating segments is predicted from adjacent segments. Then in § 3.2.2, we describe how D2L automatically constructs such generalizations.

When published as a journal article, we will make our code and data publicly available.

3.2.1 The Structure of Generalizations

D2L learns a local rule, from which the surface form of alternating segments (the rule's target) can be predicted from left-adjacent or right-adjacent segments after a (possibly empty) set of segments has been deleted. Consequently, we characterize a generalization as the composition of a local rule with the a tier projection (21). The tier $T \subseteq \Sigma$ is a subset of the segment inventory Σ, and $x = x_i, \ldots, x_n$ is an input sequence.

(21) $g(x) = r_{\text{adj}} \circ \text{proj}(x, T)$

Clearly, the proj$(\cdot, T)$ can interchangeably be viewed as creating a new sequence in which only segments in T are preserved, or as creating a new sequence in which all segments not in T are deleted. Consequently, it will sometimes be useful to refer to the complement of T as the **deletion set** $D \triangleq \Sigma \setminus T$. Thus, tier projection is an *erasing function* (Heinz et al., 2011), which creates a new sequence $t_1, \ldots, t_m$, where every segment $x_i \notin T$ is deleted. The tier-segments are annotated with their indices in the original sequence so that the results of the rule application can be written to the correct position in the output sequence. For instance, a [+sib] tier projection is shown in (22) for a hypothetical input. We use '<·>' to denote a tier projection. The superscripts denote the annotation of each sibilant's position in the input sequence.

(22) proj(/ʃoku-s-is/, [+sib]) = <ʃ$^{(1)}$s$^{(5)}$s$^{(7)}$>

Once the tier is projected, the local rule r_{adj} applies over it. This r_{adj} has the structure of an SPE[2] (Chomsky and Halle, 1968) rule with either a left (23a) or a right context (23b).

(23) a. $A \rightarrow B$ / C __

b. $A \rightarrow B$ / __ C

In this chapter, $A \rightarrow B$ can either be Agree$(A, \mathcal{F})$, which sets A's features $f \in \mathcal{F}$ to match those in C, or Disagree$(A, \mathcal{F})$, which sets them to the opposite of C's. This follows Nevins (2010) in drawing analogy to syntactic Agree (Chomsky, 2001b,a). We discuss this point further in the dissertation's conclusion (§ 7.3.7). This brings the rule schemas to those in (24).

(24) a. Agree$(A, \mathcal{F})$ / C __
Agree$(A, \mathcal{F})$ / __ C

b. Disagree$(A, \mathcal{F})$ / C __
Disagree$(A, \mathcal{F})$ / __ C

An example sibilant harmony rule is (25).

(25) $g(x) = \text{Agree}([\text{+sib}], \{\text{ant}\})/[\text{+sib}]\_\_ \circ \text{proj}(x, [\text{+sib}])$

Because harmony patterns tend to spread (Nevins, 2010; Burness et al., 2021), the rules apply iteratively. If the rule has a left-context, it applies left-to-right; otherwise it applies right-to-left. If one wishes to learn non-iterative processes, the model can straight-forwardly be extended by applying rules simultaneously instead of iteratively.

The output sequence is generated by copying every non-tier element to the output directly, and copying every tier element to the output position corresponding to its annotated index. Example

[2]SPE is a common acronym for the title of Chomsky and Halle (1968), *The Sound Pattern of English*.

(26) shows the sibilant rule (25) applying to a hypothetical input—underlines show the rule applications. The annotated indexes are a shorthand for the graph structure, demonstrated in (3) of the dissertation's introduction, which is implied by the indexes.

(26) /ʃoku-s-is/ → <ʃ$^{(1)}$s$^{(5)}$s$^{(7)}$> → <$\underline{\text{ʃ}^{(1)}\text{ʃ}^{(5)}}s^{(7)}$> → <ʃ$^{(1)}$$\underline{\text{ʃ}^{(5)}\text{ʃ}^{(7)}}$> → [ʃokuʃiʃ]

The restriction of operations to AGREE and DISAGREE and the context to either a single position to the left or right is to keep the model and its discussion succinct. To handle epenthesis, deletion, and contexts larger than a single segment (e.g., intervocalic voicing), D2L could use the more general rule structures of PLP in § 5.

For the interested reader, the appendix includes discussions of connections with search-and-copy accounts of vowel harmony (§ A.1.1) and what D2L's generalizations correspond to in formal-language-theoretic accounts of phonology (§ A.1.2). § 7.3.7 discusses possible connections to syntactic Agree.

3.2.1.1 Examples and Expressivity

These tier-based generalizations can express a wide range of behaviors, including *transparency, parasitic harmony*, and *blocking*.

Transparent segments do not participate in an alternation. For example, as described in § 3.1, the Finnish vowels {i, e} are transparent in vowel harmony (19). This can be expressed by excluding the neutral vowels from the tier T = [−cons] \ {i, e}. Since all tier vowels harmonize, and r_{adj} applies over the tier projection, the context can be C = [−cons]. This is stated in (27), where A = [?back] targets vowels unspecified for [back], causing them to harmonize with a vowel to the left on the tier, which includes all vowels except {i, e}.

(27) AGREE([?back], {back})/[−cons] __ ∘ proj(·, [−cons] \ {i, e})

For example, if the ESS suffix vowel in [k$\underline{\text{o}}$tin$\underline{\text{ɑ}}$] 'home'-ESS from (19b) is underlyingly underspecified for [back], rule (27) would derive the surface form as in (28). Here /A/ is the /−round,+low,?back/ vowel, which alternates between [+back] [ɑ] and [−back] [æ].

(28) /koti-nA/ → <o$^{(2)}$A$^{(6)}$> → <$\underline{\text{o}^{(2)}\text{ɑ}^{(6)}}$> → [k$\underline{\text{o}}tin\underline{\text{ɑ}}$]

In *parasitic harmony*, segments only harmonize with respect to some feature when they agree in another feature. In Kachin Khakass, only [+high] {i, ɯ, y, u} suffixal vowels undergo rounding harmony, and only with [+high] stem vowels (29a). They fail to harmonize with [−high] {e, ø, a, o} stem vowels (29b) and [−high] suffixal vowels never harmonize (29c). The following examples come from Korn (1969) and Burness et al. (2021, p. 18).

(29) a. [kyn-ny] ‘day-ACC’
[kuʃ-tun] ‘of the bird’
b. [ok-tɯn] ‘of the arrow’
c. [kyn-ge] ‘to the day’
[pol-za] ‘if he is’

This can be captured by including only [+high] vowels on the tier: T = [+high, −cons] as in (30). The context can again be [−cons] since the tier already excludes [−high] vowels.

(30) AGREE([?round], {round})/[−cons] __ ∘ proj(·, [+high, −cons])

In some cases, such as Turkish secondary rounding harmony (see § 3.5.1), [+high] vowels undergo rounding harmony with vowels of any height. This can be expressed by including all vowels on the tier, while excluding [−high] vowels from the target (31).[3]

(31) AGREE([−high, ?round], {round})/[−cons] __ ∘ proj(·, [−cons])

In some cases, harmony or dissimilation is *blocked* by some segments. For example, in Khalkha Mongolian (Nevins, 2010, p. 137) (32) the rounding harmony in (32a) is blocked by the [+round] vowels {u, ʊ} (32b).

(32) a. [tor-oːd] ‘be.born-PERF’
[ɔr-ɔːd] ‘enter-PERF’
b. [tor-uːl-eːd] ‘be.born-CAUS-PERF’
[ɔr-ʊːl-aːd] ‘enter-CAUS-PERF’

Critically, the PERF affix vowels {e, a} in (32b) are [−round], implying that they do not harmonize opaquely with the [+round] {u, ʊ} blockers. This blocking can be expressed by including [+round] on the tier, but excluding [+round] from the context (33).

(33) AGREE([?round], {round})/[−round] __ ∘ proj(·, [−cons])

This must be combined with a default [−round] value to account for [−round] vowels surfacing in (32b) when harmony is blocked. We discuss discovering defaults in § 3.2.2.3.

[3]Alternatively, if only alternating vowels are underspecified, [−high] vowels would be fully specified underlyingly and thus excluded automatically by the [?round] condition.

3.2.2 Learning

D2L follows the steps in (34). The cases of assimilation and dissimilation are symmetrical. For clarity, we describe D2L in the context of assimilation, and discuss in § 3.2.2.4 how D2L automatically infers whether the alternation is assimilatory or dissimilatory.

(34) **Input:** (UR, SR) pairs V and a set of segments A that alternate on features $\mathcal{F}$

1. Initialize tier $T = \Sigma$ (equivalently deletion set $D = \emptyset$)
2. **While** $T \neq \emptyset$ **do**
3. – $g_l = \textsc{Agree}(A, \mathcal{F})/C_l\_\_ \circ \text{proj}(\cdot, T)$
4. – $g_r = \textsc{Agree}(A, \mathcal{F})/\_\_C_r \circ \text{proj}(\cdot, T)$
5. – $g = \arg\max_{g \in \{g_l, g_r\}} \text{acc}(g, V)$ (§ 3.2.2.1)
6. – **If** $\text{sat}(g, V)$ **then** (§ 3.2.2.1)
7. — **Return** g
8. – Remove from T segments adj. to A on T that cannot account for the alternation (§ 3.2.2.2)

The input to D2L is a set of input-output (UR, SR) pairs, such as (35a), where we use /S/ for alternating sibilants.[4] The alternating segments, A, and what features they alternate on, $\mathcal{F}$, can be directly computed from discrepancies between these inputs and outputs (35b). In (35a), since sometimes /S/ → [ʃ] and sometimes /S/ → [s], 'S' ∈ A and 'ant' ∈ $\mathcal{F}$. In our presentation of D2L, we treat alternating segments as underlyingly underspecified (e.g. /S/). This is likely not crucial to D2L, which can also operate over fully specified default forms (e.g. /s/). Chapter § 4 provides a hypothesized account of how underlying forms are constructed.

We will use (35) as a toy example throughout our presentation of D2L. The (UR, SR) pairs are (35a) and the alternating segments and features are (35b).

(35) a. /ʃoku-S-iS/ → [ʃokuʃiʃ]
/apʃa-S/ → [apʃaʃ]
/ʃun-iS/ → [ʃuniʃ]
/soki-S/ → [sokis]
/simo-S-iS/ → [simosis]
/ut-S/ → [uts]

b. $A = \{\text{S}\}, \mathcal{F} = \{\text{ant}\}$

Initially no segments are deleted, so $T = \Sigma$ (34; step 1). After initialization, D2L enters the while-loop (34; step 2) and constructs left and right rules (34; step 3)-(34; step 4). To do so, the

[4]The default form of /S/ is [s]; we discuss how D2L discovers this in § 3.2.2.3.

sets C_l/C_r are constructed to contain every segment tier-adjacent (on the left for g_l and the right for g_r) to an alternating segment. For the words in (35), the first left and right rules are those in (36) because the segments to the left of /S/ are {u, i, a, o, t} and to the right of /S/ are {i, ⋉}, where '⋉' denotes a right word boundary.

(36) Tier, deletion set, and rules at the first iteration
$T = \Sigma$, $D = \emptyset$
g_l = AGREE({S}, {ant})/{u, i, a, o, t}__ ∘ proj(·, Σ)
g_r = AGREE({S}, {ant})/__{i, ⋉} ∘ proj(·, Σ)

The accuracy of these rules is then computed (34; step 5) and the more accurate rule is checked to see if it is a sufficiently good generalization (34; step 6)

3.2.2.1 Computing Accuracy and Rule Quality

The accuracy of a rule g is straight-forwardly defined (37c) as the number of correct predictions made by g over the training instances V (37b) divided by its total number of predictions over the training instances (37a).

(37) a. $n(g, V) \triangleq$ number of g's predictions over V
b. $c(g, V) \triangleq$ number of g's correct predictions over V
c. $\text{acc}(g, V) \triangleq \frac{c(g,V)}{n(g,V)}$

Since rules are applied iteratively, the number of applications and correct applications are computed iteratively as well. Thus, the sibilant harmony rule (38a), which states that underlying /S/ should agree in anteriority with the preceding sibilant (after projecting a [+sib] tier), has two applications (both correct) over the input (38b). We denote rule applications with underlines.

(38) a. AGREE({S}, {ant})/[+sib]__ ∘ proj(x, [+sib])
b. /ʃoku-S-iS/ → <ʃ$^{(1)}$S$^{(5)}$S$^{(7)}$> → <$\underline{\text{ʃ}^{(1)}\text{ʃ}^{(5)}}S^{(7)}$> → <ʃ$^{(1)}\underline{\text{ʃ}^{(5)}\text{ʃ}^{(7)}}$> → [ʃokuʃiʃ]
(+1) **(+1)**

An application is considered incorrect if either (a) it predicts the incorrect surface form, or (b) the target cannot AGREE/DISAGREE with the contextual segment due to the relevant feature being unspecified. For example, consider rule (39a), which states that underlying /S/ should agree in anteriority with the preceding consonant. This rule will make a correct prediction for /apʃa-S/ → [apʃaʃ] in (39b), since the underlying /S/ taking its anteriority from /ʃ/ indeed leads to it correctly surfacing as [ʃ]. However, for /ʃun-iS/ → [ʃuniʃ], /S/ harmonizing with [+ant] [n] will lead to incorrect surface form [+ant] [s] (39c).

(39) a. AGREE({S}, {ant})/[+cons]__ ∘ proj(·, [+cons])
b. /apʃa-S/ → <p$^{(2)}$ʃ$^{(3)}$S$^{(5)}$> → <p$^{(2)}\underline{\text{ʃ}^{(3)}\text{ʃ}^{(5)}}$> → [apʃaʃ] ✓
c. /ʃun-iS/ → <ʃ$^{(1)}$n$^{(3)}$S$^{(5)}$> → <ʃ$^{(1)}\underline{\text{n}^{(3)}\text{s}^{(5)}}$> → [ʃunis] ✗

A rule like (40a), which states that /S/ should agree in anteriority with the preceding vowel (all segments Σ projected), will produce an error on /apʃa-S/ → *[apʃaS] for the latter reason: /S/ cannot take the feature value for [ant] from /a/, assuming vowels are not specified for consonantal features.

(40) a. AGREE({S}, {ant})/[−cons]__ ∘ proj(·, Σ)
b. /apʃa-S/ → <a$^{(1)}$p$^{(2)}$ʃ$^{(3)}$a$^{(4)}$S$^{(5)}$> → <a$^{(1)}$p$^{(2)}$ʃ$^{(3)}\underline{\text{a}^{(4)}\text{S}^{(5)}}$> → [apʃaS] ✗

Thus, for vocabulary (35), rule (39a) makes a correct prediction for only the sibilants preceded by a consonant that correctly predicts the surface form for /S/. This is shown in (41), where [+ant] = {n, t, d, s} and [−ant] = {b, p, m, k, g, ʃ} ('*' marks errors).

(41) /ʃoku-S-iS/ → <ʃ$^{(1)}$k$^{(3)}$S$^{(5)}$S$^{(7)}$> → <ʃ$^{(1)}\underline{\text{k}^{(3)}\text{ʃ}^{(5)}}S^{(7)}$> → <ʃ$^{(1)}k^{(3)}\underline{\text{ʃ}^{(5)}\text{ʃ}^{(7)}}$>
→ [ʃokuʃiʃ]
/apʃa-S/ → <p$^{(2)}$ʃ$^{(3)}$S$^{(5)}$> → <p$^{(2)}\underline{\text{ʃ}^{(3)}\text{ʃ}^{(5)}}$> → [apʃaʃ]
/ʃun-iS/ → <ʃ$^{(1)}$n$^{(3)}$S$^{(5)}$> → <ʃ$^{(1)}\underline{\text{n}^{(3)}\text{s}^{(5)}}$> → [ʃuni*s]
/soki-S/ → <s$^{(1)}$k$^{(3)}$S$^{(5)}$> → <s$^{(1)}\underline{\text{k}^{(3)}\text{ʃ}^{(5)}}$> → [soki*ʃ]
/simo-S-iS/ → <s$^{(1)}$m$^{(3)}$S$^{(5)}$S$^{(7)}$> → <s$^{(1)}\underline{\text{m}^{(3)}\text{ʃ}^{(5)}}S^{(7)}$> → <s$^{(1)}m^{(3)}\underline{\text{ʃ}^{(5)}\text{ʃ}^{(7)}}$>
→ [simo*ʃi*ʃ]
/ut-S/ → <t$^{(2)}$S$^{(3)}$> → <$\underline{\text{t}^{(2)}\text{s}^{(3)}}$> → [uts]

There are 8 instances of /S/, and (39a) predicted the correct surface form for 4 of these. Thus, $n(g, V) = 8$, $c(g, V) = 4$, and $\text{acc}(g, V) = 4/8 = 1/2$.

The function $\text{sat}(g, V)$ is a boolean function that returns 'True' iff g is satisfactorily accurate over the training data V. We use the Tolerance Principle of Yang (2016) as the criterion due to its cognitive basis and prior success in computational modeling, lexical, and experimental studies, reviewed in § 2.5. Recall that the Tolerance Principle hypothesizes that learners accept a linguistic generalization when it is cognitively more efficient to do so, and provides a quantitative, categorical threshold for this tipping point in terms of the generalization's scope and how many exceptions it has (see Yang 2016, ch. 3 for the threshold's derivation). When the threshold is met and a generalization accepted, the exceptions to the generalization can be lexicalized. In the current work, using the Tolerance Principle for evaluating generalizations is achieved—in terms of (37)—via (42).

(42)
$$\text{sat}(g, V) \triangleq n(g, V) - c(g, V) \leq \frac{n(g, V)}{\ln n(g, V)}$$

For the initial rules (36), both g_l and g_r make $n(g_l, V) = n(g_r, V) = 8$ predictions over the vocabulary (35), because there are 8 underlying /S/'s. The left rule in (36) makes only 1 correct prediction: /unt-S/ → [unts]; the remaining 7 predictions error because the underlying /S/'s cannot harmonize with adjacent vowels. The right rule in (36) makes no correct predictions. Thus, $\text{acc}(g_l, V) = 1/8$ and $\text{acc}(g_r, V) = 0/8$. Since the former is more accurate, it is chosen (34; step 5). However, since $8 - 1 > 8/\ln 8$ (i.e., $7 > 3.85$), $\text{sat}(g, V)$ = 'False' at (34; step 6), and the tier must be updated.

3.2.2.2 Updating the Tier

In order for the rule to apply to them, the alternating segments A must always be preserved on the tier. Moreover, any segment currently tier-adjacent to an alternating segment from which the correct surface form cannot be computed cannot be on the tier if the alternation is to be predictable from adjacent dependencies. Consequently, these unuseful segments, which are present in the context sets C_l and C_r, are added to the deletion set D (43).

(43) $D \cup \{s \in C_l \cup C_r :$ agreeing with s is not possible or yields the wrong surface form$\}$

In our example, $\{\text{u, i, a, o}\} \cup \{\text{i}, \ltimes\}$ are added to D. The segment $\{\text{t}\}$ is not added because agreeing /S/ to /t/ correctly yields [s]. For segments that do not occur tier-adjacent to an alternating segment (e.g., /k/) or yield the correct surface form (e.g., [+ant] /t/ adjacent to an /S/ that surfaces as [+ant] [s]), no conclusion can be drawn about whether they should be on the tier or off it. Thus, D2L takes the smallest natural class that contains all of D but none of A, and removes this from the tier (44) at (34; step 8).

(44) $T \setminus \arg\min_{\{\text{nat class } N : D \subseteq N \wedge A \cap N = \emptyset\}} |N|$

The arg min ranges over natural classes,[5] each of which we refer to with N. The condition $D \subseteq N$ requires that all segments to be deleted (D) are included in the natural class (N), so that they are excluded from the tier. The condition $A \cap N = \emptyset$ requires that no alternating segments (A) are included in N, so that they are preserved on the tier. The arg min returns the smallest natural class satisfying these conditions ($|N|$ is the size of N).

If no such natural class exists, it removes D verbatim, allowing for idiosyncratic tiers that do not fit neatly into a natural class. In our example, both [+cons] and [+sib] include $A = \{\text{S}\}$ and exclude $D = \{\text{u, i, a, o}, \ltimes\}$, so N in (44) ranges over their complements [−cons] and [−sib]. The class [−cons] is smaller (only the vowels) than [−sib] (both vowels and non-sibilant consonants), so

[5]For ease of exposition, we consider only natural classes describable with a single feature, but clearly the same process could apply over natural classes requiring more features to specify.

[−cons] is deleted; equivalently T = [+cons]. This yields the tier projections shown in (45c), from which the rules (45d) are derived in the second iteration of the while loop.

(45) Deletion set, tier, tier projections, and rules at the second iteration

a. $D = \{\text{u, i, a, o, }\ltimes\}$

b. T = [+cons]

c. /ʃoku-S-iS/ → <$\text{ʃ}^{(1)}\text{k}^{(3)}\text{S}^{(5)}\text{S}^{(7)}$>
/apʃa-S/ → <$\text{p}^{(2)}\text{ʃ}^{(3)}\text{S}^{(5)}$>
/ʃun-iS/ → <$\text{ʃ}^{(1)}\text{n}^{(3)}\text{S}^{(5)}$>
/soki-S/ → <$\text{s}^{(1)}\text{k}^{(3)}\text{S}^{(5)}$>
/simo-S-iS/ → <$\text{s}^{(1)}\text{m}^{(3)}\text{S}^{(5)}\text{S}^{(7)}$>
/ut-S/ → <$\text{t}^{(2)}\text{S}^{(3)}$>

d. g_l = Agree({S}, {ant})/{k, S, ʃ, n, m, t}__ ∘ proj(·, [+cons])
g_r = Agree({S}, {ant})/__{S, ⋉} ∘ proj(·, [+cons])

The sets $C_l = \{\text{k, S, ʃ, n, m, t}\}$ and $C_r = \{\text{S, }\ltimes\}$ are computed from the segments to the left and right of the alternating /S/ on the [+cons] tier. The new rules g_l and g_r again apply to all 8 /S/ segments, so $n(g_l, V) = n(g_r, V) = 8$. The right rule g_r makes 0 correct predictions because /S/ cannot harmonize with '⋉' or a right-adjacent /S/ that also is not specified for anteriority. The left rule makes 4 correct predictions, as shown in (46): 2 on the first word, 1 on the second, and 1 on the last word ('*' marks errors).

(46) /ʃoku-S-iS/ → <$\text{ʃ}^{(1)}\text{k}^{(3)}\text{S}^{(5)}\text{S}^{(7)}$> → <$\text{ʃ}^{(1)}\underline{\text{k}^{(3)}\text{ʃ}^{(5)}}\text{S}^{(7)}$> → <$\text{ʃ}^{(1)}\text{k}^{(3)}\underline{\text{ʃ}^{(5)}\text{ʃ}^{(7)}}$>
→ [ʃokuʃiʃ]
/apʃa-S/ → <$\text{p}^{(2)}\text{ʃ}^{(3)}\text{S}^{(5)}$> → <$\text{p}^{(2)}\underline{\text{ʃ}^{(3)}\text{ʃ}^{(5)}}$> → [apʃaʃ]
/ʃun-iS/ → <$\text{ʃ}^{(1)}\text{n}^{(3)}\text{S}^{(5)}$> → <$\text{ʃ}^{(1)}\underline{\text{n}^{(3)}\text{s}^{(5)}}$> → [ʃuni*s]
/soki-S/ → <$\text{s}^{(1)}\text{k}^{(3)}\text{S}^{(5)}$> → <$\text{s}^{(1)}\underline{\text{k}^{(3)}\text{ʃ}^{(5)}}$> → [soki*ʃ]
/simo-S-iS/ → <$\text{s}^{(1)}\text{m}^{(3)}\text{S}^{(5)}\text{S}^{(7)}$> → <$\text{s}^{(1)}\underline{\text{m}^{(3)}\text{ʃ}^{(5)}}\text{S}^{(7)}$> → <$\text{s}^{(1)}\text{m}^{(3)}\underline{\text{ʃ}^{(5)}\text{ʃ}^{(7)}}$>
→ [simo*ʃi*ʃ]
/ut-S/ → <$\text{t}^{(2)}\text{S}^{(3)}$> → <$\underline{\text{t}^{(2)}\text{s}^{(3)}}$> → [uts]

The successes are when the tier-adjacent consonants {k, ʃ, t} match the surface anteriority of /S/, and the errors are when {n, k, m} do not. Since $4 > 8/\ln(8)$, the rule is still not sufficiently accurate, according to the Tolerance Principle threshold (42). Since the segments {n, k, m} led to incorrect predictions, they are added to D, yielding $D = \{\text{u, i, a, o, n, k, m, }\ltimes\}$. The natural class [+cons] no longer separates $A = \{\text{S}\}$ from $D = \{\text{u, i, a, o, n, k, m, }\ltimes\}$, because D now contains vowels and consonants. However, [+sib] does separate A and D, so T = [+sib] is set at (34; step 8), yielding (47).

(47) Deletion set and tier at the third iteration

D = {u, i, a, o, n, k, m, ⋉}

T = [+sib]

At the third iteration, the new left rule (48a), which projects a [+sib] tier, correctly predicts all the surface forms (48b) except for /ut-S/, where it fails to apply because there is no stem sibilant (i.e., the rule underextends). When this happens, D2L attempts to infer a default form for the alternating segment, as discussed in the next section (§ 3.2.2.3). The right rule g_r will have zero accuracy for the same reason as the prior iteration, so the left rule is evaluated under the Tolerance Principle at (34; step 6). Since g_l applies to only the first 7 instances of /S/ (the 8th being handled by the default case, as we discuss next), $n(g_l, V) = 7$. As shown in (48b), the rule predicts the correct surface form in all 7 cases, so $c(g_l, V) = 7$. Since $0 \leq 7/\ln 7$, this rule is accepted and returned (34; step 7).

(48) Successful rule and its predictions at the third iteration

a. g_l = AGREE({S}, {ant})/{s, ʃ}__ ∘ proj(·, [+sib])

b. /ʃoku-S-iS/ → <ʃ$^{(1)}$S$^{(5)}$S$^{(7)}$> → <$\underline{\text{ʃ}^{(1)}\text{ʃ}^{(5)}}S^{(7)}$> → <$\underline{\text{ʃ}^{(1)}\text{ʃ}^{(5)}}\text{ʃ}^{(7)}$> → [ʃokuʃiʃ]

/apʃa-S/ → <ʃ$^{(3)}$S$^{(5)}$> → <$\underline{\text{ʃ}^{(3)}\text{ʃ}^{(5)}}$> → [apʃaʃ]

/ʃun-iS/ → <ʃ$^{(1)}$S$^{(5)}$> → <$\underline{\text{ʃ}^{(1)}\text{ʃ}^{(5)}}$> → [ʃuniʃ]

/soki-S/ → <s$^{(1)}$S$^{(5)}$> → <$\underline{\text{s}^{(1)}\text{s}^{(5)}}$> → [sokis]

/simo-S-iS/ → <s$^{(1)}$S$^{(5)}$S$^{(7)}$> → <$\underline{\text{s}^{(1)}\text{s}^{(5)}}S^{(7)}$> → <s$^{(1)}\underline{\text{s}^{(5)}\text{s}^{(7)}}$> → [simosis]

/ut-S/ → <S$^{(3)}$> → [ut*S]

3.2.2.3 Default Values

When a candidate rule underextends, D2L takes the set of alternating segments that the rule does not account for and computes the set of surface realizations of those underextensions. If they do not alternate, the surface form is taken as the default. For instance, the rule (48a) underextends for the affix sibilant in /ut-S/, which surfaces as [s] (48b). Taking [s] as the default form for /S/ works, since there are no underextensions of (48a) where /S/ surfaces as anything else. If there were such underextensions, D2L would reject the candidate rule and continue the while-loop.

As another example, when Finnish stems contain only neutral vowels, alternating affix vowels are usually [−back] by default (Ringen and Heinämäki, 1999). For example, the essive affixal vowel, which alternated between [+back] [ɑ] and [−back] [æ] in (19) surfaces as [−back] [æ] when the stem contains only the neutral vowel [e] (49).

(49) [velje-næ] road-ESS (Nevins, 2010, p. 76)

A [back] harmony rule like (27), which excludes neutral vowels {i, e} from the tier, will underextend to words like (49) with only neutral vowels. However, because these underextensions

consistently surface as [−back] vowels, D2L infers this as the default.

3.2.2.4 Assimilation vs. Dissimilation

From observing a segment that alternates, a learner will not immediately know whether the alternation is due to assimilation or dissimilation. However, attempting to account for an assimilatory alternation by dissimilating from something in the phonological environment (or vice versa) is highly unlikely to yield a productive generalization. Thus, figuring out whether an alternation is assimilatory or dissimlatory should not present a serious challenge to learning. Consequently, we run the D2L algorithm (34) twice in parallel—one searching for an assimilatory rule and one searching for a dissimlatory rule. When searching for an assimilatory rule, g_l and g_r are constructed with AGREE, and when searching for dissimlatory rule, they are constructed with DISAGREE. All other aspects are identical. In most conceivable cases, only one search will yield a productive generalization, in which case the generalization from the successful search is chosen. In the unlikely case that both searches yield a generalization, the more accurate one is chosen. We never observed this scenario in any of our experiments.

The reason it is necessary to take this approach, instead of expanding (34; step 3)-(34; step 4) to include two more rules (i.e., left and right dissimilatory rules), is that it is not possible to maintain a single deletion set for both assimilation and dissimilation. In assimilation, the segments to delete are those that assimilating with yields the wrong surface form; in dissimilation they are those that dissimilating from yields the wrong surface form. Running D2L twice in parallel allows for maintaining these two, distinct deletion sets.

3.2.3 Strict Locality as a Special Case

Since D2L starts with an empty deletion set, a strictly-local alternation—one determined by string-adjacency—will be discovered on the first iteration of the algorithm. Consequently, D2L is a unified model for learning both local and non-local alternations. This fact has been recognized in tier-based accounts of non-locality (e.g., McMullin 2016).

3.3 Prior Models

Prior models for learning non-local phonological generalizations have mostly focused on phonotactics. In contrast, D2L is focused on alternations. We group prior models into statistical, § 3.3.1, formal-language-theoretic § 3.3.2, and neural network § 3.3.3 models.

3.3.1 Statistical Models

Hayes and Wilson (2008) demonstrated that an inductive baseline phonotactic learner sensitive to only fixed-length sequences failed to learn phonotactic constraints for long-distance patterns. If provided a projection of the data onto a relevant tier, the model was then able to learn relevant phonotactic constraints on that tier. However, the model did not learn what tier to project. Gouskova and Gallagher (2020) extended the Hayes and Wilson (2008) model to automatically learn projections. The authors observed that many non-local dependencies, despite being arbitrarily far away in principle, often occur within a window of three segments (i.e., a trigram). Their model uses Hayes and Wilson (2008) to extract baseline phonotactic constraints. Some of these constraints are trigram constraints of the form *X[]Y, where X and Y are sets of segments, and [] allows any segment to intervene. Gouskova and Gallagher (2020)'s model then uses these trigram constraints to project a tier, over which additional constraints are learned. The tier it constructs is the smallest natural class that contains the segments from both X and Y, so that both X and Y are preserved on the tier. For example for the data in (35a)—surface forms reproduced in (50)—if the absence of non-harmonizing sibilant trigrams is sufficiently statistically conspicuous, then the Hayes and Wilson (2008) model may learn the constraints *[s][][ʃ] and *[ʃ][][s].

(50) [ʃokuʃiʃ]
[apʃaʃ]
[ʃuniʃ]
[sokis]
[simosis]
[uts]

The smallest natural class containing both [s] and [ʃ] is [+sib], so Gouskova and Gallagher (2020)'s model would then project the [+sib] tier and re-apply Hayes and Wilson (2008) over that projection. The success of this model depends upon the trigram restriction being frequent enough in the data to discover the relevant projection, and upon the effectiveness of Hayes and Wilson (2008) at discovering constraints over the projection. Moreover, the authors recognized a limitation: the model's inability to capture more complex phenomena like opaque or blocking segments, which must be included on the tier but do not participate in the restriction. This is because the model constructs the tier based on the sets X and Y, which do not contain information about blocking segments.

Goldsmith and Riggle (2012) proposed an information theoretic model for justifying a tier-based descriptive account of Finnish vowel harmony. They used Goldsmith and Xanthos (2009)'s Hidden Markov Model (HMM) approach to extract two classes of segments that maximizes the probability of the data. Because the segments of a word tend to alternate between consonants and vowels (e.g.

CVCV is much more frequent than CCVV), this HMM approach is best-suited for extracting the two categories *consonant* and *vowel*. Goldsmith and Riggle (2012) used a Boltzmann model to score phonological ill-formedness in terms of unigram probabilities, and bigram mutual information over the surface string and the vowel tier. The intuition for the model is that it combines the frequency of segments (unigrams) with the frequency of bigrams over the words and vowel-tier projections to score phonological ill-formedness. This model is largely limited to interactions between all vowels or all consonants, because the HMM usually constructs the consonant and vowel categories. Thus, on (50), the model could find dependencies between sibilants on the consonant tier, but non-sibilant consonants would prevent some sibilants from occurring withing the purview of the bigram mutual-information computation.

3.3.2 Formal-Language-Theoretic Approaches

Heinz (2010) proposed a model for learning long-distance phonotactics based on the *precedence* relation. However some long-distance patterns cannot be accounted for in terms of the precedence relation (Heinz, 2010; Jardine and Heinz, 2016). Consequently, later formal-language-theoretic approaches have instead targeted the class of Tier Strictly-Local (TSL) constraints Heinz et al. (2011), which have been argued to subsume most or all non-local consonant interactions (McMullin, 2016). Their functional analogue, the TSL functions, are argued to cover a broad range of non-local processes (Burness et al., 2021).

In particular, Jardine and Heinz (2016) proposed a model for learning Tier-based Strictly 2-Local (TSL_2) formal languages, which are languages where the words of the language can be distinguished from the non-grammatical by a tier-sequence of length 2. The model is provably capable of learning such languages, in the sense of Gold (1967). Jardine and McMullin (2017) extended the model to handle arbitrary values of k. Jardine (2016b) applied Jardine and Heinz (2016)'s model to idealized natural language data, showing that the model successfully learns phonotactic restrictions, but only in an idealized setting where exceptions were removed and segments were pre-organized into natural classes. While most work has focused on learning classes of stringsets for long-distance phonotactics, Burness and McMullin (2019) proposed a model for learning Tier-based Strictly 2-Local *functions*, which are appropriate for long-distance phonological processes. Formal-language-theoretic models are introduced for the purposes of theoretical learnability proofs. They are intended as a starting-point for what a learning algorithm for a particular formal-language class (e.g. TSL) must look like, and are not necessarily intended for use on natural language data.

3.3.3 Neural Network Models

There have been several attempts to model aspects of vowel harmony with recurrent neural networks (RNNs). Hare (1990) proposed using an RNN to model Hungarian vowel harmony, training it on synthetic bit-sequences and finding that its assimilatory behavior on these bit-sequences mirrored some of the complexities of Hungarian vowel harmony. Rodd (1997), while not targeting vowel harmony directly, found that RNN models could use distributional information to learn phonological categories from a small Turkish corpus, and that they could learn to treat [+back] and [−back] vowels differently.

These early models were limited in their empirical scope, but in the decades since, research on neural networks (NNs) has accelerated. NNs have been used for morphological reinflection (Cotterell et al., 2016) and to revisit connectionism in the 'past-tense debate' of English morphology (Kirov and Cotterell, 2018). While NNs as cognitive models of morphophonological learning has been questioned (e.g., McCurdy et al. 2020; Belth et al. 2021) and these more recent models have not directly been applied to modeling long-distance alternations, some of the languages included in the SIGMORPHON reinflection task involve non-local dependencies. Moreover, Smith et al. (2021) used an RNN-based model to evaluate an articulatory account of height harmony in Nzebi by training the model on simulated speech to map segments to vocal-tract articulator movements.

Because prior works' problem settings have varied, it is difficult to draw conclusions about RNNs' ability to model non-local alternations, but an RNN is certainly applicable as a comparison model for D2L (see § 3.4.1.1 and § A.2 for details of how we use it as such).

3.4 Comparison to Human Behavior

3.4.1 Model Behavior on Finley (2011)

We first compare to Finley (2011)'s artificial language experiment, which we describe here. Participants were presented with training data consisting of <stem, suffixed> pairs where the suffix contained a sibilant that harmonized with a stem sibilant across a single intervening vowel (51).

(51) /di͟so-͟su/ → [di͟so͟su]
/ne͟si-͟su/ → [ne͟si͟su]
/pi͟ʃa-͟su/ → [pi͟ʃa͟ʃu]
/ku͟ʃo-͟su/ → [ku͟ʃo͟ʃu]

To make the sibilants adjacent on a tier, the vowels must be excluded, suggesting [+cons] is the relevant tier. This predicts that when vowels and non-sibilant consonants intervene, the non-sibilant consonants will block the harmony. This prediction is borne out. After training, participants were

evaluated in a two-alternative forced choice (2AFC) paradigm, where they were presented with a stem and two choices for its suffixed from: one harmonizing and one not. Learners generalized to novel instances like the training instances (52a), demonstrating that they learned a harmony pattern. However, they showed no preference for harmony in novel cases where non-sibilant consonants also intervened (52b), choosing the non-harmonizing option as often as a control group did.

(52) a. /baso-su/ → ✓ [basosu], *[basoʃu]
/deʃe-su/ → ✓ [deʃeʃu], *[deʃesu]
b. /ʃeta-su/ → ? [ʃetaʃu], ? [ʃetasu]
/ʃomi-su/ → ? [ʃomiʃu], ? [ʃomisu]

In contrast, a second experiment presented participants with training data where sibilants harmonize across both intervening vowels and non-sibilant consonants (53).

(53) /suge-su/ → [sugesu]
/sone-su/ → [sonesu]
/ʃupe-su/ → [ʃupeʃu]
/ʃako-su/ → [ʃakoʃu]

In this case, the [+sib] tier is needed in order for the sibilants to be adjacent, in which case learners should generalize to cases where only vowels intervene. The prediction is again borne out. Learners generalized to novel train-like instances (54a), and instance where only a vowel intervened (54b).

(54) a. /ʃika-su/ → ✓ [ʃikaʃu], *[ʃikasu]
/ʃege-su/ → ✓ [ʃegeʃu], *[ʃegesu]
b. /keʃu-su/ → ✓ [keʃuʃu], *[keʃusu]
/niʃa-su/ → ✓ [niʃaʃu], *[niʃasu]

3.4.1.1 Comparison Models

GR is Goldsmith and Riggle (2012)'s phonotactic model (see also § 3.3 and § 3.6). To compare the relative well-formadness of candidates, we use the Boltzmann score from pg. 882 of their paper. Following the authors, we used Laplace smoothing with 0.5 smoothing-factor.

GG is Gouskova and Gallagher (2020)'s phonotactic model (see § 3.3 and § 3.6 for further description). We use the author's code[6] and default parameters.

TSLIA is Jardine and Heinz (2016); Jardine and McMullin (2017)'s formal-language-theoretic model. We used the implementation from Aksënova (2020). This model is binary: it accepts or

[6] https://github.com/gouskova/inductive_projection_learner

rejects a candidate string. When more than one candidate is accepted by the model, we choose one at random as the best candidate.

LSTM is a Recurrent Neural Network (RNN) sequence-to-sequence model. We trained the model to predict the surface form of each underlyingly underspecified segment. This simplifies the learning problem by not requiring the model to predict the surface form for the entire sequence, but makes for a fair comparison to D2L, which also has access to the underlying forms. We used a Pytorch (Paszke et al., 2019) implementation, and discuss architecture specifics, training procedure, and hyperparameter tuning in § A.2.

3G is a trigram phonotactic model, which assigns a probability to each candidate in terms of the trigrams it contains—how frequent they are in the training data. As in GR, we used Laplace smoothing with a smoothing-factor of 0.5.

3.4.1.2 Setup

We use the training and test items recorded in Finley (2011, p. 15)'s appendix. Each model is trained on the training instances, then probed to choose between each pair of items in the 2AFC test set. We treat the sibilant in [-su]/[-ʃu] as underlyingly /S/, unspecified for [ant]. Using default /s/ instead has no meaningful impact on the results. For the comparison phonotactic models, we treat the item to which the model assigns the higher score as its choice. If the model assigns the same score to both items, then the choice is made at random. Since D2L produces an output for an input, we use its produced form as its choice if the produced form matches one of the 2AFC choices. Otherwise the choice is made at random. To simulate multiple participants, we run each model 30 times and report averages and standard deviations. This is important because some randomness is introduced due to GR, GG, and 3G being stochastic, and because when a model assigns the same score to both 2AFC choices, the choice is made at random. Following Finley (2011), each of the 24 training items appears 5 times in the total exposure set, and the items are presented in a random order for each of the 30 runs. We list the segment features in Tab. A.1 in the appendix.

3.4.1.3 Results

The results for the first experiment are given in Tab. 3.1, and for the second experiment in Tab. 3.2. The tables report, for each model, the fraction of the test instances where the model picked the harmonizing choice. The HUM row records a '✓' whenever Finley (2011) reported that the experimental-group participants chose the harmonizing choice significantly more often than the control-group participants. It records an '✗' wherever the two groups chose the harmonizing choice

Table 3.1: Results for Finley (2011)'s first experiment, where training instances involved sibilants harmonizing across intervening vowels. Test instances are of three types: train (Old), novel train-like (New Train-Like), and novel items where both vowels and non-sibilant consonants intervene between sibilants (Novel). D2L generalizes in exactly the cases where humans do, and does not generalize in exactly the cases where humans do not.

	Train CV$\underline{S}$V-$\underline{S}$V		
	CV$\underline{S}$V-$\underline{S}$V (Old)	CV$\underline{S}$V-$\underline{S}$V (New Train-Like)	$\underline{S}$VCV-$\underline{S}$V (Novel)
HUM	✓	✓	✗
D2L	1.0000 ± 0.00	1.0000 ± 0.00	0.4694 ± 0.15
GR	0.6222 ± 0.33	0.7424 ± 0.23	0.5722 ± 0.08
GG	0.4611 ± 0.13	0.5091 ± 0.14	0.5000 ± 0.13
TSLIA	0.5556 ± 0.15	0.5424 ± 0.14	0.5222 ± 0.16
LSTM	0.8833 ± 0.20	0.7909 ± 0.18	0.8528 ± 0.26
3G	1.0000 ± 0.00	1.0000 ± 0.00	0.7028 ± 0.04

at statistically indistinguishable rates.[7] If, over a set of test items, a model makes the harmonizing choice significantly more often than a control model that makes a random selection from the two choices, then we treat the model as having generalized to those test items. The test of significance is made with a one-sided t-test that compares the average model performance over 30 runs to that of the random control model. The null hypothesis is that the tested model's average performance is equal to the random control model's. We use Welch's t-test, which does not assume equal variance, and a significance level of $\alpha = 0.99$. We shade a cell gray if the model matches the human result—i.e. iff either both HUM and the model generalized or neither generalized.

3.4.1.4 Discussion

D2L matches the human results in all cases. When trained on CV$\underline{S}$V-$\underline{S}$V words, D2L learns a [+cons] tier (55a), which generalizes harmony to novel CV$\underline{S}$V-$\underline{S}$V (second column, Tab. 3.1), but not $\underline{S}$VCV-$\underline{S}$V words (third column, Tab. 3.1). In contrast, when trained on $\underline{S}$VCV-$\underline{S}$V words, D2L learns a [+sib] tier (55b), which generalizes harmony to both novel $\underline{S}$VCV-$\underline{S}$V words (second column, Tab. 3.2) and novel CV$\underline{S}$V-$\underline{S}$V words (third column, Tab. 3.2).

(55) a. AGREE({S}, {ant})/{s, ʃ} __ ∘ proj(·, [+cons])

b. AGREE({S}, {ant})/{s, ʃ} __ ∘ proj(·, [+sib])

D2L has non-zero variance in the third column of Tab. 3.1 because rule (55a) does not extend to

[7] We chose to report the human results in this way because Finley (2011) only reported these conclusions, not the actual mean rates of the harmonizing choice.

Table 3.2: Results for Finley (2011)'s second experiment, where training instances involved sibilants harmonizing across both vowels and non-sibilant consonants. Test items are of three types: train (Old), novel train-like (New Train-Like) and novel items where only vowels intervene between sibilants (Novel). Both humans and D2L learned a harmony pattern (Old) and extend it to both New Train-Like and Novel test instances.

	Train SVCV-SV		
	SVCV-SV (Old)	SVCV-SV (New Train-Like)	CVSV-SV (Novel)
HUM	✓	✓	✓
D2L	1.0000 ± 0.00	1.0000 ± 0.00	1.0000 ± 0.00
GR	0.9889 ± 0.06	0.4139 ± 0.01	0.2472 ± 0.01
GG	0.5222 ± 0.13	0.5222 ± 0.13	0.5139 ± 0.13
TSLIA	0.5639 ± 0.10	0.5111 ± 0.14	0.5222 ± 0.16
LSTM	0.9000 ± 0.20	0.9278 ± 0.25	0.7361 ± 0.33
3G	1.0000 ± 0.00	0.5833 ± 0.00	0.2500 ± 0.00

the SVCV-SV test items, so the choice between [-su]/[-ʃu] was made randomly for each.[8] Moreover, D2L's performance is error-free over a type of test item when its generalization extends to it because, as a computational model, it is not subject to the experimental complexities applicable to human participants, who likely have imperfect memory of the exposure data and imperfect attention during the experiment.

The comparison models are largely ineffective at generalizing at all from the limited training data, with the exception of LSTM. However, when trained on CVSV-SV words, LSTM generalizes to both CVSV-SV and SVCV-SV test words, whereas humans do not generalize to SVCV-SV words (Tab. 3.1). Thus, LSTM fails to exhibit the blocking behavior of non-sibilant consonants in the first experiment.

GR is able to achieve moderate performance generalizing to CVSV-SV words when trained on words of the same type (second column, Tab. 3.1). For SVCV-SV words, GR expected harmony at a rate slightly above chance. Despite failing the statistical test, it does come close to matching human performance because the intervening C prevents the sibilants from being adjacent on the consonant tier that the Hidden Markov Model learns. However, when trained on SVCV-SV, no interactions between sibilants are visible to the model and it is unable to generalize beyond the training data (Tab. 3.2).

GG appears unable to generalize from these small datasets. While this may seem to be a limita-

[8]Had we treated the underlying suffixal sibilant as default /s/, when the generalization does not apply, D2L would always choose [-su]. While this would not lead to variance, it would still perform no better than chance, since no more than 50% of the test items take the [-su] suffix for the harmonizing choice. Thus, the treatment of the underlying sibilant is not consequential.

tion of the data, note that humans were able to generalize from the same training data. In the next section, we turn to McMullin and Hansson (2019)'s study, which involves substantially more data, and is thus more instructive regarding GG's generalization behavior. Similarly, TSLIA is unable to generalize from the training data.

3G frequently chooses the harmonizing form for SVCV-SV words when trained on CVSV-SV words, and rarely chooses the harmonizing form for CVSV-SV words when trained on SVCV-SV words. This is the opposite of what humans did.

3.4.2 Model Behavior on McMullin and Hansson (2019)

McMullin and Hansson (2019) replicated results similar to Finley (2011)'s, extending them to both liquid harmony and dissimilation. In these experiments the harmony/dissimilation was *regressive*. Participants were first exposed to a practice phase where they were presented with verb stems followed by a past-tense form, which added a [-ɹu] suffix to the stem, and the same stems followed by a future-tense form, which added a [-li] suffix to the stem. These practice-phase stems did not contain liquids: they allowed the participants to learn the relevant morphology.

Next, in the training phase, participants were presented with <stem, past, future> triplets. The two training settings of Finley (2011) where doubled by McMullin and Hansson (2019), performing each experiment in both assimilatory and dissimilatory settings. In all experiments, the data contained 50% distractors, which were stems with no liquid.

In what the authors labeled experiment 1a, the stem liquid harmonized regressively with the affix liquid across an intervening vowel (56) (treating the alternating stem-liquid as underlyingly underspecified /L/).

(56) /toboLe-ɹu/ → [toboɹeɹu]
/toboLe-li/ → [toboleli]
/dumiLi-ɹu/ → [dumiɹiɹu]
/dumiLi-li/ → [dumilili]

Symmetrically, in experiment 2a, the stem liquids *dissimilated* regressively from the affix liquid across an intervening vowel (57).

(57) /toboLe-ɹu/ → [toboleɹu]
/toboLe-li/ → [toboɹeli]
/dumiLi-ɹu/ → [dumiliɹu]
/dumiLi-li/ → [dumiɹili]

Like Finley (2011)'s study, participants generalized the training pattern to novel words of the same CVCVLV-LV form, but did not extend it to CVLVCV-LV or LVCVCV-LV forms, where

the liquids crossed more than just vowels. This is the predicted behavior if participants construct a [+cons] tier.

In experiments 1b and 2b, the participants were presented with assimilatory (58a) and dissimilatory (58b) instances of the form CVL̲VCV-L̲V.

(58) a. /teL̲omu-ɹ̲u/ → [teɹ̲omuɹ̲u]
/teL̲omu-l̲i/ → [tel̲omul̲i]
/poL̲eku-ɹ̲u/ → [poɹ̲ekuɹ̲u]
/poL̲eku-l̲i/ → [pol̲ekul̲i]

b. /teL̲omu-ɹ̲u/ → [tel̲omuɹ̲u]
/teL̲omu-l̲i/ → [teɹ̲omul̲i]
/poL̲eku-ɹ̲u/ → [pol̲ekuɹ̲u]
/poL̲eku-l̲i/ → [poɹ̲ekul̲i]

A [+liquid] tier is needed for the training liquids to be adjacent, predicting that learners will extend the pattern to CVCVL̲V-L̲V and L̲VCVCV-L̲V forms. These predictions were borne out.

3.4.2.1 Setup

We follow McMullin and Hansson (2019, sec. 2)'s setup to produce the stimuli. The setup details are the same as in the prior experiment (§ 3.4.1.2) except that, following McMullin and Hansson (2019), the stimuli are only presented once each. The stimuli include the practice phase forms. We treat alternating stem liquids [l]/[ɹ] as underlyingly underspecified /L/. Treating them as default /l/ has no meaningful impact on the results. We use the same comparison models as in the prior experiment § 3.4.1.1 and list the segment features in Tab. A.2 in the appendix.

3.4.2.2 Results

The results for the experiments with CVCVL̲V-L̲V exposure data are given in Tab. 3.3, and the results for the experiments with CVL̲VCV-L̲V exposure data in Tab. 3.4. For the assimilation experiments, the tables report, for each model, the fraction of the test instances where the model picked the harmonizing choice. For the dissimilatory experiments, they report the fraction where the disharmonizing choice was made. The HUM row records a '✓' whenever McMullin and Hansson (2019) reported that the experimental-group participants extended the training pattern to test instances of the form in the corresponding column, and an '✗' where they did not. Gray cells mark where the models match the human result. As before, the models are compared to a control model that makes a random selection in the 2AFC test.

Table 3.3: Results from McMullin and Hansson (2019)'s Experiment 1a (assimilation) and 2a (dissimilation), where training instances involved liquids interacting across intervening vowels. D2L matches human behavior in all cases.

	Train CVCVLV-LV (assimilation)			Train CVCVLV-LV (dissimilation)		
	CVCVLV-LV	CVLVCV-LV	LVCVCV-LV	CVCVLV-LV	CVLVCV-LV	LVCVCV-LV
HUM	✓	✗	✗	✓	✗	✗
D2L	1.000 ± 0.00	0.480 ± 0.06	0.497 ± 0.08	1.000 ± 0.00	0.520 ± 0.06	0.503 ± 0.08
GR	0.987 ± 0.07	0.640 ± 0.08	0.495 ± 0.08	1.000 ± 0.00	0.363 ± 0.08	0.509 ± 0.08
GG	0.925 ± 0.19	0.507 ± 0.04	0.504 ± 0.02	0.985 ± 0.05	0.552 ± 0.10	0.607 ± 0.04
TSLIA	0.477 ± 0.09	0.526 ± 0.09	0.495 ± 0.11	0.523 ± 0.09	0.474 ± 0.09	0.505 ± 0.11
LSTM	0.862 ± 0.22	0.859 ± 0.22	0.822 ± 0.22	0.850 ± 0.23	0.848 ± 0.23	0.830 ± 0.23
3G	0.969 ± 0.00	0.639 ± 0.07	0.506 ± 0.08	1.000 ± 0.00	0.362 ± 0.07	0.494 ± 0.08

Table 3.4: Results from McMullin and Hansson (2019)'s. Experiment 1b (assimilation) and 2b (dissimilation), where training instances involved liquids interacting across intervening vowels. D2L matches human behavior in all cases.

	Train CVLVCV-LV (assimilation)			Train CVLVCV-LV (dissimilation)		
	CVCVLV-LV	CVLVCV-LV	LVCVCV-LV	CVCVLV-LV	CVLVCV-LV	LVCVCV-LV
HUM	✓	✓	✓	✓	✓	✓
D2L	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.00	1.000 ± 0.00
GR	0.423 ± 0.05	0.643 ± 0.08	0.509 ± 0.04	0.577 ± 0.05	0.357 ± 0.08	0.491 ± 0.04
GG	0.500 ± 0.04	0.475 ± 0.07	0.496 ± 0.04	0.933 ± 0.12	0.922 ± 0.13	0.935 ± 0.11
TSLIA	0.477 ± 0.09	0.526 ± 0.09	0.495 ± 0.11	0.523 ± 0.09	0.474 ± 0.09	0.505 ± 0.11
LSTM	0.745 ± 0.25	0.748 ± 0.25	0.739 ± 0.25	0.833 ± 0.24	0.833 ± 0.24	0.831 ± 0.23
3G	0.401 ± 0.06	0.642 ± 0.07	0.497 ± 0.11	0.599 ± 0.06	0.358 ± 0.07	0.503 ± 0.11

3.4.2.3 Discussion

D2L matches the human results in every setting, and is the only model to do so. When trained on CVCVL̲V-L̲V words, D2L learns (59a) and (59c), with a [+cons] tier projection. When trained on CVL̲VCV-L̲V words, D2L learns (59b) and (59d), with a {l, ɹ, L} liquid tier. These results also demonstrate that D2L is able to correctly infer whether a process is assimilatory or dissimilatory (see § 3.2.2.4).

(59) a. AGREE({L}, {ant, cor, lat})/{l, ɹ} __ ∘ proj(·, [+cons])

b. AGREE({L}, {ant, cor, lat})/{l, ɹ} __ ∘ proj(·, {l, ɹ, L})

c. DISAGREE({L}, {ant, cor, lat})/{l, ɹ} __ ∘ proj(·, [+cons])

d. DISAGREE({L}, {ant, cor, lat})/{l, ɹ} __ ∘ proj(·, {l, ɹ, L})

When trained on CVCVL̲V-L̲V words (Tab. 3.3), D2L is the only model to match human behavior in all cases, though some comparison models also come close. However, when trained on CVL̲VCV-L̲V words (Tab. 3.4), only D2L and LSTM come close to human behavior.

Similarly to the previous experiment (§ 3.4.1), when trained on CVCVL̲V-L̲V words, LSTM generalizes to words where non-liquids intervened and humans did not generalize (Tab. 3.3). These results suggest that LSTM may not be able to express blockers, instead learning a dependence between L's independent of what segments intervene.

Because the HMM of GR usually learns the consonant/vowel tiers, the interacting liquids in CVL̲VCV-L̲V words are blocked by the intervening C, which prevents the model from substantially generalizing. Similarly, the interacting liquids are beyond the trigram sensitivity of 3G, so its occasional above-chance performance is only due to coincidental statistical regularities in the exposure data unrelated to the liquid interaction. TSLIA again performs at chance in all settings.

The asymmetry of GG between assimilation and dissimilation deserves some discussion. Because no -LVL- sequences occur in CVL̲VCV-L̲V training instances, the Hayes and Wilson (2008) model extracts a trigram constraint *[l,ɹ][][l,ɹ], which GG uses to project a liquid tier. In the assimilation experiment, GG learns the single constraint *[l,ɹ][l,ɹ] on this tier, not two constraints *[lɹ] and *[ɹl]. Consequently, it cannot distinguish between assimilitory and non-assimilatory sequences. In the dissimilation experiment, GG learns the single tier-constraint *[ll], as well as a non-tier constraint *[−cor], which applies to [ɹ] but not [l]. Consequently [ɹl] » [ll] due to the tier constraint, which has a much higher weight than the non-tier constraint. Moreover, [lɹ] » [ɹɹ] because the latter violates *[−cor] twice. Thus, the two constraints coincidentally conspire to yield a functional generalization. However, the coincidental result only works for dissimilation, and the asymmetry with assimilation is not consistent with human behavior, which was similar for assimilation and dissimilation.

3.5 Learning Natural Language Alternations

3.5.1 Turkish Vowel Harmony

The Turkish vowel inventory, (60), has 4 [+back] and 4 [−back] vowels, all of which participate in [±back] harmony (Kabak, 2011, p. 2).

(60)

	front		back	
	unround	round	unround	round
high	i	y	ɯ	u
low	e	ø	ɑ	o

Affix vowels alternate between [+back] vowels when the final stem vowel is [+back] and [−back] when the final stem vowel is [−back], as shown in (61) repeated from (17).

(61) [dɑ̲l-lɑ̲r-ɯ̲n] branch-PL-GEN (Kabak, 2011, p. 3)
[je̲r-le̲r-i̲n] place-PL-GEN (Kabak, 2011, p. 3)
[i̲p-le̲r-i̲n] rope-PL-GEN (Nevins, 2010, p. 28)

In addition to back/front harmony, the [+high] vowels participate in secondary rounding harmony, as exemplified by the GEN affix (62). The [+high] vowels harmonize with the final vowel of the stem, both when the the stem vowel is [+high] (62a) and [−high] (62b).

(62) a. [i̲p-i̲n] rope-GEN (Nevins, 2010, p. 29)
[jy̲z-y̲n] face-GEN (Kabak, 2011, p. 3)
[kɯ̲z-ɯ̲n] girl-GEN (Nevins, 2010, p. 29)
[bu̲z-u̲n] ice-GEN (Kabak, 2011, p. 3)
b. [e̲l-i̲n] hand-GEN (Kabak, 2011, p. 3)
[sø̲z-y̲n] word-GEN (Nevins, 2010, p. 29)
[sɑ̲p-ɯ̲n] stalk-GEN (Nevins, 2010, p. 29)
[jo̲l-u̲n] road-GEN (Kabak, 2011, p. 3)

The question arises whether Turkish exhibits two independent harmony processes or one. We follow Nevins (2010) in treating it as one process. In this view, some affixal [+high] vowels exhibit surface alternation on both [back] and [round], whereas [−high] vowels do not. This leads to an assymetry in underlying underspecification, and the restriction of rounding harmony to only [+high] vowels is a reflection of that. Consequently, AGREE is taken to copy only the underspecified vocalic features from the closest vowel, i.e., both [back] and [round] for [?back, +high, ?round], but only [back] for [?back, ±high, ±round].

The most informative evidence towards this treatment is the above observation that [+high] vowels that alternate on roundness take their [±round] value from *any* vowel, regardless of height. This treatment allows the harmony to be stated as one generalization, but it is not critical. If we were to treat them as separate processes, or with fully-specified underlying default forms, D2L could be run twice—once for feature [±back] and once for [±round]—to construct two generalizations.

Affixes can be exceptional—not participating in vowel harmony—or even half-half harmonic, where only one of two vowels participates. However, as a result of their exceptional status, such affixes do not alternate on the surface, so there is no motivation for underspecification (Nevins, 2010, sec. 2.6).

3.5.1.1 Setup

We use data from the MorphoChallenge (Kurimo et al., 2010), which provides 1760 frequency-annotated and morphologically-segmented words. We used these as our surface forms. Underlying stem vowels match their surface forms. For each affix vowel, we checked all surface-realizations of the vowel and set any feature that alternated on the surface as underlyingly unspecified. This follows *invariant transparency*, which states that learners only posit abstract underlying forms when doing so is necessary to account for surface alternations (Kiparsky, 1968; Peperkamp et al., 2006; Ringe and Eska, 2013; Tesar, 2013; O'Hara, 2017; Richter, 2018, 2021). For instance the vowel of the PL affix [-lɑr]/[-ler] is set to /?back, −high, −round/, and the GEN affix vowel [-in]/[-yn]/[-ɯn]/[-un] is set to /?back, +high, ?round/. We skipped any word that contained an affix with less than ten occurrences. This processes yielded 1198 words. Turkish orthographic vowels map directly to phonemes. We mapped orthographic consonants to phonemes directly, except we treated <c> as [ʤ], <x> as [ks], <y> as [j], and <q> as [k]. As Caplan and Kodner (2018, p. 1442) point out, this direct grapheme-to-phoneme conversion may be noisy, but this noise can serve to test a model's robustness. We list the segment features in Tab. A.3 in the appendix.

For training data, we sampled 80% of the words (958), weighted by frequency. We used the remaining 20% of words for testing. We repeated this 30 times with a different random sample each time. We report the average performance across the 30 runs. To get a predicted surface form for the phonotoactic comparison models, we computed all possible specifications of an input word's underspecified affix vowels, presented the model with each of these candidates, and selected the candidate with the highest score. For instance, for input /dɑl-l[?back, −high, −round]r-[?back, +high, ?round]n/, the candidate surface forms are [dɑl-lɑr-in], [dɑl-lɑr-yn], [dɑl-lɑr-ɯn], [dɑl-lɑr-un], [dɑl-ler-in], [dɑl-ler-yn], [dɑl-ler-ɯn], [dɑl-ler-un]. The comparison models are the same as the prior experiments.

Table 3.5: Accuracy of models on held-out test words, when learning Turkish vowel harmony.

Model	Test Accuracy
D2L	**0.9840 ± 0.01**
GR	0.7913 ± 0.28
GG	0.8914 ± 0.07
TSLIA	0.2249 ± 0.02
LSTM	0.7249 ± 0.18
3G	0.5614 ± 0.02

3.5.1.2 Results

The results shown in Tab. 3.5 demonstrate that D2L learns a vowel harmony generalization that robustly generalizes to unseen test words. D2L's accuracy is substantially higher than the comparison models and is consistent with acquisition studies, which reveal that Turkish-speaking children as young as 2;0—when their vocabulary likely contains under a thousand words—already know vowel harmony well enough to extend it to nonce words (Altan, 2009). The reason D2L does not get 1.0 accuracy is due to the fact that the data contains some exceptions to vowel harmony,[9] which D2L is able to tolerate and still find the generalization (63). Under the Tolerance Principle, these exceptions can be lexicalized.

(63) $\textsc{Agree}([\text{?back}], \{\text{back}, \text{round}\}) / [-\text{cons}]\ \_\_ \circ \text{proj}(\cdot, [-\text{cons}])$

This rule states that vowels with neutralized backness (i.e., [?back]) take their [±back] value and, if also neutralized [?round], their [±round] value from a vowel to the left on a projected vowel tier. These are the main generalizations standardly reported in linguistic analyses (Kabak, 2011). The statement of primary [back] and secondary [round] harmony as a single generalization is possible because we followed Nevins (2010)'s treatment of them as a single process (see above discussion in the § 3.5.1 introductory description of Turkish).

3.5.2 Finnish Vowel Harmony

The Finnish vowel inventory (64) (Suomi et al. 2008, sec. 3.1) has 8 vowels, 6 of which participate in front/back harmony, as discussed in § 5.1 and repeated in (65).

[9] See chapter § 4.3 for more discussion of Turkish and its exceptions.

(64)

	front		back	
	unround	round	unround	round
high	i	y		u
mid	e	ø		o
low	æ		ɑ	

The vowels {i, e} are neutral, neither participating in nor blocking harmony (65b). When a stem contains only neutral vowels, alternating affix vowels are [−back] by default (65c).

(65)
a. [pøytæ̲-næ̲] table-ESS (Ringen and Heinämäki, 1999, p. 304)
[poutɑ̲-nɑ̲] fine weather-ESS (Ringen and Heinämäki, 1999, p. 304)
b. [kø̲ti-nɑ̲] home-ESS (Ringen and Heinämäki, 1999, p. 305)
c. [velje-næ] road-ESS (Nevins, 2010, p. 76)

Sinc the neutral vowels are [−back], the default case (65c) could instead be characterized as the affix vowel harmonizing with the neutral vowels only when no non-neutral vowels are present. However, the treatment of [−back] as a default is a preferred analysis, due to the existence of the language Uyghur, which has the same vowel organization, except that alternating vowels are default [+back] when the stem contains only [−back] neutral vowels (Lindblad 1990; Nevins 2010, p. 77-78).

3.5.2.1 Setup

Finnish data also comes from the MorphoChallenge (Kurimo et al., 2010). There are 1835 frequency-annotated words with morphological segmentations. The surface form of these affixes varied extensively in ways beyond just vowel alternations (e.g., the GEN affix has forms [n], [en], [ten], ...) making it challenging to compute precisely how vowels alternate. Consequently, we treated all affix vowels as alternating and underlyingly unspecified for feature [back]. This choice potentially overestimates the number of harmony exceptions, and thus potentially makes the problem more challenging for D2L, while not effecting the comparison phonotactic models, which learn directly from surface forms. We dropped words with only one occurrence, yielding a set of 1219 words. Finnish orthographic vowels map directly to phonemes. We mapped orthographic consonants to phonemes directly, but treated <x> as [ks] and <c>, <q> as [k]. The segment features are in Tab. A.4 in the appendix.

For training data, we again sampled 80% of words (975) weighted by frequency and used the other 20% for testing. We repeated this sampling 30 times, and report the average performance across the 30 runs. To to get a predicted surface form for the phonotoactic comparison models, we

Table 3.6: Accuracy of models on held-out test words, when learning Finnish vowel harmony.

Model	Test Accuracy
D2L	**0.9799 ± 0.01**
GR	0.8519 ± 0.03
GG	0.8092 ± 0.04
TSLIA	0.7198 ± 0.02
LSTM	0.8164 ± 0.03
3G	0.8421 ± 0.02

again computed candidate surface forms by permuting affix vowels between back and front, and select the candidate that the model assigns the highest score.

3.5.2.2 Results

The accuracies shown in Tab. 3.6 reveal that D2L is the most accurate model. D2L learns the generalization (66), where unspecified vowels take their [±back] feature from the vowel to the left, skipping consonants and neutral vowels (66a). D2L also learned that if no non-neutral vowel is to the left, the surface vowel should be [−back] (66b).

(66) a. AGREE([?back], {back})/[−cons] __ ∘ proj(·, [−cons] \ {i, e})

b. Elsewhere [−back]

The tier correctly excludes the neutral vowels while preserving all other vowels. On a small number of simulations, the tier contained one or two spurious segments—[f] and/or [b]—because they never occur in between an affix vowel and a stem vowel in the training data and, consequently, never interfere with harmony. The elsewhere condition matches the default value given in accounts of Finnish for what happens when a stem contains only neutral vowels (Ringen and Heinämäki, 1999). Note that because the segments do not fall neatly into a natural class, D2L simply listed the tier segments explicitly; we presented it here with natural classes and set operations for clarity.

3.5.3 Latin Liquid Dissimilation

In Latin, default /l/ (20a) dissimilates to [r] when preceded by /l/ across varying distances, as shown in (20b) and repeated in (67b). This process is usually discussed and most clearly seen in the adjectival *-alis*/*-aris* affix. The dissimilation of /l/ is blocked by intervening /r/ (67c). Cser (2010) argues that intervening [−cor] consonants also block dissimilation (67d).

(67)

Table 3.7: Accuracy of models on held-out test words after learning Latin liquid dissimilation.

Model	Test Accuracy
D2L	**0.9653 ± 0.03**
GR	0.5292 ± 0.09
GG	0.0431 ± 0.03
TSLIA	0.1569 ± 0.05
LSTM	0.7736 ± 0.11
3G	0.6083 ± 0.09

a. *nav-alis* ‘naval’
b. *popul-aris* ‘popular’
lun-aris ‘lunar’
c. *flor-alis* ‘floral’
d. *pluvi-alis* ‘rainy’
leg-alis ‘legal’

3.5.3.1 Setup

The data comes from the Perseus project (Smith et al., 2000), which contains Old and Classical Latin texts from the 3rd century BCE through the 2nd century CE. Words are annotated for frequency. We extracted words containing two liquids and ending in *-alis*/*-aris*, and removed non-adjectives. The result is a dataset of 121 words. We treat /-alis/ as the underlying form of both *-alis* and *-aris* words. We map orthographic segments directly to phonemes, treating <v> as the semivowel [w], <c> as [k], <x> as [ks], and <ll> as [l]. We also dropped the <h> from <th>, <kh>, and <ph>, treating <h> as marking aspiration. We list the segment features in Tab. A.5 in the appendix.

As the prior experiments, the training data is an 80% frequency-weighted sample of words (97), and the testing data is the remaining 20% of words. This sampling was repeated 30 times, and models' performances are computed as averages over these 30 train/test splits.

3.5.3.2 Results

Again (Tab. 3.7), D2L is the most accurate model by a substantial margin. D2L discovered rule (68), where /l/ dissimilates to [r] when preceded by /l/.

(68) DISAGREE({l}, {lat})/{l}__ ∘ proj(·, {l, r, k, h, f, b, p, w, m})

As Cser (2010) discusses, /r/ between the two /l/'s blocks the dissimilation. Cser (2010, p. 38) also argues that [−cor] consonants block dissimilation. D2L's generalization corroborates this

conclusion, as the tier contains the [−cor] consonants {k, h, f, b, p, w, m}. We found that [+cor] [d] is also preserved on the tier in 3 of the 30 simulations. Cser (2010) notes that there is no data either way for whether [d] blocks. Since D2L deletes only when necessary, the presence of [d] on the tier in some simulations is consistent with Cser (2010)'s analysis.

3.6 Discussion

We intend D2L as a cognitive model for how humans learn phonological alternations, at roughly the *algorithmic level* in the sense of Marr (1982). That is, D2L constitutes a hypothesis that human learners roughly follow the steps laid out by D2L—tracking adjacent dependencies and iteratively discarding those that do not lead to successful generalization. This chapter constitutes a presentation of that hypothesis (§ 3.2) and initial evidence for it on the grounds of prior artificial language studies (§ 3.4) and new computational modeling experiments (§ 3.5). As an explicit computational model, it can be used to generate predictions for further experimental studies. We believe D2L is of importance to phonological theory because it provides an account of how phonological tiers are not only learnable by tracking only adjacent dependencies, but that such representations arise naturally from a learner that tracks adjacency first (Saffran et al., 1996, 1997; Aslin et al., 1998; Santelmann and Jusczyk, 1998; Gómez, 2002; Newport and Aslin, 2004; Gómez and Maye, 2005) and iteratively removes adjacent dependencies that are unhelpful to generalization. Prior accounts of how the mind may construct phonological tiers have been scant at best.

When Goldsmith (1976) and others introduced autosegmental tiers into phonological theory, the representation was quickly put to effective use. It has been used to describe complex phenomena involving apparently non-local dependencies, like vowel harmony (Clements, 1980; Goldsmith, 1985) and non-concatinative morphology (McCarthy, 1981). Moreover, autosegmental representations have been part of a larger movement to provide representations that render long-distance dependencies as local, including metrical stress patterns (Liberman and Prince, 1977) and syntactic relations (Chomsky, 2001a,b).

Given the abstract nature of these representations, they are, perhaps by definition, not explicitly present in surface structures. Consequently, attention has rightly been given to justifying an appeal to such representations in linguistic theory. For instance, Hayes and Wilson (2008) argued for their legitimacy in terms of an inductive baseline: a model without a phonological tier or metrical grid—the inductive baseline—failed to learn non-local phonotactic dependencies, but was more successful when provided them. Goldsmith and Riggle (2012) argued that a vowel tier is a legitimate representation for Finnish vowel harmony on the grounds that an information-theoretic model can describe a Finnish lexicon using fewer bits if it tracks bigram probabilities over a vowel tier than if it does not. Computational analysis of phonology has suggested that most of (at least segmen-

tal) phonology can be characterized in terms of adjacent dependencies on some tier representation (Heinz et al., 2011; McMullin, 2016; Burness et al., 2021). Even strictly local dependencies—e.g. the voicing assimilation in (89)—can be characterized as the special case where the relevant tier contains all segments (i.e. the deletion set is empty). This computational analysis has produced theoretical learning results that demonstrate the strong learning-theoretic benefits of restricting search to tier-local generalizations, but these models are not intended to describe algorithmically how humans learn phonology. Similarly, Hayes and Wilson (2008) and Goldsmith and Riggle (2012) were attempts to justify the use of autosegmental tier representations in phonological theory, and were not intended as accounts of how children may construct these representations during acquisition.

Prior to the present work, Gouskova and Gallagher (2020)'s model (GG) comes the closest to providing a possible account for how tiers may be constructed. However, we believe that model is not fully adequate for multiple reasons. First, GG deviates in fundamental ways from human behavior in the artificial language experiments investigated in § 3.4. Second, the authors recognized that their model is ill-suited for handling opaque or blocking segments, as was seen in the case of Latin liquid dissimilation § 3.5.3. The dissimilation is blocked by [−cor] consonants, which GG's inability to express contributes to its poor performance (6.13% accuracy generalizing to held-out test words) compared to D2L (95.96% accuracy) and even the simple trigram (3G) baseline (35.29% accuracy). Third, the GG model is always looking for the possibility of tier-based generalizations in tirgram constraints, which differs from D2L, where tiers are the natural consequence of adjacent dependencies being inadequate. This difference is critical, as the artificial language experiment in chapter § 6 suggests that learners do not track non-adjacent dependencies when adjacent dependencies suffice.

Consequently, we believe D2L is, at present, the best hypothesis about how the mind may construct phonological tiers. Next we discuss which aspects of D2L lead it to perform more successfully in the experiments than comparison models (§ 3.6.1), followed by current limitations of D2L and future directions (§ 3.6.2).

3.6.1 D2L vs. Other Models

The key to D2L's success is its iterative creation of the deletion set, the complement of which constitutes a tier. It is D2L's tracking of only adjacent dependencies that leads it to add segments that do not work to the deletion set. Consequently, it is—somewhat ironically—D2L's proclivity for adjacency that drives its learning of non-local alternations. We will highlight why this leads it to match human behavior in the artificial language experiments (§ 3.4) and succeed at learning natural language alternations (§ 3.5) even when they contain complexities like blocking segments (§ 3.5.3).

The artificial language experiments (§ 3.4) used poverty-of-stimulus paradigms, which precisely design the exposure (training) data so as to underdetermine what process underlies the observed alternation. Thus, when Finley (2011)'s exposure data contains CVSV-SV words, it could be that harmony applies strictly to sibilants across intervening vowels (i.e. SVS) or also across intervening non-sibilants (e.g. SVCVSV). The exposure data contains no words where sibilants are separated by more than a vowel, so the exposure data underdetermines which of these generalizations underlies the harmony. Similarly, when the exposure data contains SVCV-SV words, it could be that harmony applies strictly to sibilants where both consonants and vowels intervene (but not when *only* a vowel intervenes) or whenever anything other than a sibilant intervenes. Again, the choice is underdetermined. The same logic holds for McMullin and Hansson (2019)'s experiments.

As evidenced by the human behvior, something is critically different between the cases where the exposure data contains CVSV-SV words compared to when it contains SVCV-SV words. D2L's incremental deleting of adjacent dependencies leads it to delete only vowels (V) when exposed to CVSV-SV words, and both vowels and non-sibilant consonants (V and C) when exposed SVCV-SV words, mirroring the assymetry observed in human behavior. In contrast, consider the way GG constructs a tier. When the exposure data is CVSV-SV words and SVCV-SV words, the Hayes and Wilson (2008) model, which GG uses, may observe that non-harmonizing SVS sequences are conspicuously absent and learn the constraints *[s][][ʃ] and *[ʃ][][s]. When the exposure data is SVCV-SV words, it may notice that *any* (harmonizing or not) SVS sequences are conspicuously absent and learn the constraints *[s][][ʃ], *[s][][s], *[ʃ][][s], and *[ʃ][][ʃ].[10] In both cases, because GG uses the smallest natural class containing X and Y in *X[]Y constraints, it will project the smallest natural class containing {s, ʃ}, which is the [+sib] tier. Thus, fundamentally, GG is not capable of distinguishing the case where sibilants harmonize only across intervening vowels from the case where sibilants harmonize across both vowels and non-sibilant consonants.

When we turn to natural language data, D2L is able to handle blockers for the same reason. Consider the Latin liquid dissimilation from § 3.5.3. For words where the dissimilation of /l/ to [r] is blocked by an intervening /r/ or [−cor] consonants, the /l/ simply surfaces faithfully as [l], and thus does not require that {l} or [−cor] be added to the deletion set. Consequently, they are preserved on the tier and correctly block dissimilation. If we treat the alternating *-alis* liquid as underlyingly underspecified /L/ instead of default /l/, the story does not change much. Both [r] and the [−cor] consonants are [−lat], while [l] is [+lat]. Consequently, after deleting vowels from the tier, any /r/ or [−cor] consonants tier-adjacent to /L/ will lead it to correctly (albeit opaquely) dissimilate to [+lat] [l], and the blockers will be perserved on the tier.

In the same scenario, Hayes and Wilson (2008) model, used by GG, may induce the constraint

[10] In practice, we found that the exposure dataset was small enough (24 items) that Hayes and Wilson (2008) did not consistently pick up on these regularities.

*[l][][l] if such sequences are sufficiently absent in the training data. But again, GG induces the smallest natural class containing {l}, which would be [+lat], since [l] is the only lateral consonant. As a result, neither the [r] nor [−cor] blockers will be projected on the tier. In other words, GG is only sensitive to the harmonizing or dissimilating segments and thus cannot capture blockers unless they coincidentally fall in the smallest natural class subsuming X and Y in the *X[]Y constraints that Hayes and Wilson (2008)'s model extracts.

Goldsmith and Riggle (2012)'s model (GR), which was only intended as an information-theoretic attempt to justify the use of a vowel tier in analyzing Finnish vowel harmony, used Goldsmith and Xanthos (2009)'s HMM to extract tiers. However, this approach is only well-suited for extracting the consonant and vowel tiers. This is because Goldsmith and Xanthos (2009)'s HMM is a state-transition model with two hidden states; it extracts the two classes of segments that maximizes the probability of the data. Intuitively, this is maximized when the two hidden categories transition sequentially in words. In such a case, the probability of transitioning from one state to the next will be high and the probability of staying in the current state will be low. In most imaginable linguistic cases, words tend to switch between consonants and vowels—for example CVCVCV and CVCCVC are more frequent than CCVV or CCCCVCC. This is hardly a categorical fact, but it is a robust statistical trend. Thus, the HMM's two hidden states tend to robustly correspond to the categories *consonant* and *vowel*. For this reason, GR is not able to flexibly extend to alternations involving tiers other than the [+cons] or [+vowel] tiers.

The Jardine and McMullin (2017) model (TSLIA) was proposed primarily for proving theoretical learnability results about TSL stringsets. Such proofs require certain assumptions about the data available to the learner. These conditions specify a *characteristic sample*, which, as discussed by Jardine and McMullin (2017, p. 75), may not be satisfied in natural language data (e.g., due to interaction with other phonotactic constraints). TSLIA's performance suggests that such a sample is indeed not present in our datasets.

Perhaps the best performing comparison model was the LSTM neural network model. While D2L achieves much higher accuracy in the natural language experiments, the LSTM was the only model to achieve at least 0.7 accuracy on all three natural language experiments. Moreover, LSTM generalized harmony/dissimilation in the artificial language experiments. However, the LSTM fails to match human behavior, as it generalizes harmony/dissimilation to all types of test instances, including those where harmony is blocked by intervening segments. This suggests that LSTM may not be well-suited for capturing blockers.

3.6.2 Future Directions

D2L is a unified model of learning both string-local and tier-local dependencies, because the learner starts with an empty deletion set. Consequently, local interactions like the English plural alternation (16) and non-local interactions like the Turkish plural alternation (17) can be learned by the same model. In this chapter, we considered single assimilatory and dissimilatory processes. The work in § 5 and § 4 provides a broader framework for how D2L could fit in a larger theory of phonological learning that includes epenthesis, deletion, and multiple generalizations.

In its current form, D2L learns categorical rules. Many phonological processes are variable, so extending D2L to handle variation is an important step for future research. This will likely involve the current component, which constructs a tier and a local rule over it, but would need to allow the rules to be probabilistic.

We believe another promising line for future research involves investigating whether metrical stress and tonal patterns can be learned by a similar process. For example, Jardine (2016a) argued that while most of segmental phonology can be characterized in terms of tier-locality, there are numerous processes in tonal phonology (e.g. tonal plateauing) that may require greater expressive power (unbounded circumambient) to account for. Moreover, McCollum et al. (2020) argued that Tutrugbu ATR harmony requires the same expressive power as Jardine (2016a)'s tonal analyses, suggesting that unbounded circumambient processes may be more prevalent in segmental phonology than previously thought. The key characteristic of unbounded circumambient phenomena is the involvement of dependencies that are arbitrarily far away in *both directions*. Thus, future research could investigate whether D2L's approach of iteratively deleting adjacent, unuseful, segments could render local the relevant dependencies from both directions.

CHAPTER 4

Towards an Algorithmic Account of Underlying Forms

Some of the material in this chapter has been accepted at the 2023 Society for Computational in Linguistics *(SCiL) conference (Belth, 2023c).*

A traditional concept in phonological theory is that of the underlying form. However, the history of phonology has witnessed a debate about how abstract underlying representations ought to be allowed to be, and a number of arguments have been given that phonology should abandon such representations altogether. In this chapter, we consider an algorithmic, learning-based approach to the question. We propose a model that, by default, constructs concrete representations of morphemes. When and only when such concrete representations make it challenging to generalize in the face of the sparse statistical profile of language, our proposed model constructs abstract underlying forms that allow for effective generalization. We consider the highly agglutinative language, Turkish, and the heavily-studied Dutch noun voicing alternation as two case studies. We demonstrate that the underlying forms that our model constructs account for the complexities of Turkish phonology resulting from its multifaceted vowel harmony, and enable the highly-accurate prediction of novel surface forms, demonstrating the importance of some underlying forms to generalization. We then argue that the model provides a possible learning-based account of why Dutch-learning children do not productively extend voicing alternations: such alternations are not prevalent enough in a child's input to prevent the construction of productive and accurate morphophonological generalizations from concrete underlying forms.

4.1 Introduction

A traditional conception of phonological theory involves abstract underlying representations (URs) together with phonological processes (stated as rules or constraints) mapping between this abstract level of representation and a concrete, surface-level representation. Debates in the 1960's and

1970's questioned how abstract URs should be allowed to be (Hyman, 2018, p. 597), with a particularly famous article by Kiparsky (1968) arguing that the positing of non-concrete representations should only be done when motivated. Any perception of this debate as fading in subsequent years is probably better attributed to the field moving on to other questions than it is to a satisfactory resolution of the debate (Anderson, 2021).

Indeed, some phonologists have taken the position that URs should not be used in phonological theory because doing so is "(i) wrong, (ii) redundant, (iii) indeterminate, (iv) insufficient, or (v) uninteresting," as Hyman (2018, p. 591) summarized the objections. Meanwhile, much of the work on learning phonology has either focused on surface restrictions (e.g., Hayes and Wilson 2008) or continued to assume URs (e.g., Tesar and Smolensky 1998; Boersma 1997), abstracting away from the question of how (and if) such representations are constructed (see Jarosz 2019 for a summary).

One of the main justifications for the use of underlying representations is to capture generalizations. For example, the form of the English plural affix—[z], [s], or [əz]—depends on the stem-final segment, but is predictable from the stem-final segment, as in (69).

(69) [dɑg-z]
[kæt-s]
[hɔrs-əz]

Positing an underlying /-z/ derived by process into [z], [s], or [əz] allows this generalization to be captured. However such an analysis is not strictly necessary. The allomorphs could each be listed along with a set of sounds each occurs after, or the apparent relationship between singulars and plurals could be ignored altogether and both forms could simply be memorized.

In some cases, experimental and acquisition evidence suggests that learners do have productive, rule-like knowledge. For instance, children overextend morphophonological generalizations (e.g., MacWhinney 1978) and apply them to nonce words (e.g., Berko 1958). But in other cases, experimental and acquisition evidence suggests that learners do not have productive, rule-like knowledge. Some Dutch nouns, for instance, alternate in form between the singular and the plural. Stem-final obstruents are syllabified as a coda in the singular, but as the onset of the plural [-ən] suffix in the plural. Due to devoicing of obstruents in final position, this sometimes leads to stem-final obstruents being voiced in the plural and voiceless in the singular, as in (70a), while others are voiceless throughout the paradigm, as in (70b).

(70) a. [bɛt] 'bed' [bɛdən] 'bed'-PL
b. [pɛt] 'cap' [pɛtən] 'cap'-PL

The underlying form of alternating Dutch nouns is often interpreted as having an underlyingly voiced obstruent that is neutralized in final position; for instance, /bɛd/ is taken as the underlying

form of 'bed,' which is transformed into [bɛt] by a productive devoicing rule. However, acquisition and experimental evidence shows no evidence of Dutch-learning children extending the voicing alternation productively (Zamuner et al., 2006; Kerkhoff, 2007; Zamuner et al., 2012). For instance, children have difficulty recognizing or producing [slɑt] as the singular form of the plural nonce word [slɑdən].[1]

How then are we to choose from the possible analyses of (69)? Is the desire to capture a generalization sufficient motivation to choose the /-z/ analysis? In this chapter we propose a learning-based approach to this question. Specifically, we propose a computational model that assumes, by default, that underlying forms are fully concrete. The model attempts to form morphological generalizations out of sheer necessity to deal with the sparse statistical profile of language (Yang 2016, ch. 2; Chan 2008).[2] The question then becomes learning-based: when does surface-alternation of a morpheme prevent the learner from forming morphological generalizations from concrete representations? In some—but critically not all—cases, surface-alternations are pervasive enough to drive the learner to resort to abstract URs in order to effectively generalize. We present the model in § 4.2.

We evaluate the model on natural-language corpuses of the highly agglutinative language Turkish, demonstrating both when abstract URs are necessary for generalization and when they are not (§ 4.3). When combined with the model from Chapter 3 for learning local and non-local alternations, the proposed model achieves high accuracy generalizing to held-out test words (§ 4.3.4). We then evaluate the model on a natural-language corpus of Dutch nouns (§ 4.4). The alternating nouns are not prevalent enough to drive our model to construct abstract URs, and thus allows for highly accurate generalization to held-out test words without a productive generalization for the alternation. This provides a possible learning-based account of the fact that Dutch-learning children show no evidence of having productive knowledge of the voicing alternation.

4.2 Model

4.2.1 Model Input

The input to the model is a set of morphologically-analyzed surface forms. An example input of nine forms is shown in Tab. 4.1. These word forms are processed by the model incrementally, modeling the growth of a learner's lexicon.

While morphological segmentation is an important area of study in its own right, we believe it is a justified assumption given experimental evidence that infants can effectively morphologically segment nonce words. These results have been observed for French-learning 11mo-old (Marquis

[1] See § 4.4.1 for a detailed discussion.

[2] See § 2.2 for a discussion of sparsity.

Surface Form	Morphological Analysis
1. [buz-lɑr]	'ice-PL'
2. [kɯz-lɑr]	'girl'-PL
3. [el-ler]	'hand-PL'
4. [jer-ler-in]	'place'-PL-GEN
5. [søz-ler]	'word-PL'
6. [dɑl-lɑr-ɯn]	'branch'-PL-GEN
7. [sɑp-lɑr]	'stalk-PL'
8. [jyz-yn]	'face'-GEN
9. [ip-ler-in]	'rope'-PL-GEN

Table 4.1: An example Turkish input consisting of morphologically-analyzed surface forms.

and Shi, 2012) and English-learning 15mo-old (Mintz, 2013) infants. The finding is corroborated by results for 15mo Hungarian-learning infants, despite the high-level of agglutination in Hungarian (Ladányi et al., 2020).

4.2.2 Model Output

The output of the model is a lexicon, which contains a representation for each morpheme, and a lexicalized list of any input word forms not decomposable into those morphemes. The representation of a morpheme may be concrete or abstract, but we will refer to the representation constructed in the lexicon as a UR, regardless of its abstractness. We treat surface and underlying representations, whether concrete or abstract, as sequences of segments, where each segment is a set of distinctive features.

As discussed by Ettlinger (2008, sec. 4.3.4), the term *abstract* is not always used consistently. A UR is sometimes called *abstract* because it lacks the phonetic detail of an actual speech sound (e.g., /D/ as an alveolar stop lacking a voicing specification), or because it contains different segments from a surface form. Following the discussion in § 2.1, it is the creation of non-concrete underlying forms that leads to discrepancies between underlying and surface forms and thus necessitates the construction of a generalization (rule or constraint) to non-trivially map between these levels of representation. We thus consider a UR *concrete* if all word forms where the morpheme surfaces as something other than the UR are lexicalized. Complementarily, we consider a UR *abstract* if its surface realization must be inferred from a generalization (rule or constraint). For the time being, our treatment does not differentiate degrees of abstractness. For example, our use of *abstract* includes both /d/ derived into [t] by devoicing generalization /d/ → [t] / [−voi] __ and /D/ derived into [d] / [t] via a voice assimilation generalization, while /D/ would usually be thought of as *more* abstract than /d/ if [d] sometimes surfaces while /D/ never does. Future work will consider the

question of degrees of abstractness.

We assume, following prior work (§ 4.5), that each morpheme has a single UR. Future work will consider scenarios where this may not be the case. Future work will also consider what changes are necessary to handle nonconcatenative morphology.

4.2.3 Model Description

By default, the model creates a concrete UR for each morpheme. Prior work (§ 4.5) often resorts to phonological processes to produce the various surface forms of a morpheme at the first instance of surface alternation. Our model differs from this approach by treating underlying forms as concrete even after the first instance of surface alternation. Instead of immediately collapsing surface forms into a single, abstract UR, our model simply lexicalizes all word forms in which a morpheme occurs as something other than its most frequent form. It is only when the resulting lexicalization becomes unsustainable (see § 4.2.4) that the model then constructs abstract underlying forms from which the surface realizations are derived by morphophonological process.

The pseudocode for the algorithm is shown in (71).[3] As discussed in § 4.2.1, the input to the model is an incremental stream of morphologically-analyzed surface forms. Whenever the model receives a new surface form (71; step 1), it initially creates a concrete underlying form for each morpheme, storing the most frequent[4] form of the morpheme concretely (71; step 3), and lexicalizes any wordforms that contain a different form of the morpheme (91; step 8). However, if too many wordforms in the lexicon are exceptions—where the measurement of "too many" occurs as described in § 4.2.4—the model instead constructs an abstract UR (91; step 5) and then learns a phonological process, via a separate model (see § 4.2.6), to account for the resulting alternation.

(71) **Input:** Incremental stream of morphologically analyzed SRs

1. **While** surface form in input **do**
2. – **For** morpheme in segmentation **do**
3. — Morpheme UR ← most freq form
4. — **If** too many alternative forms **do**
5. —— Construct abstract UR
6. —— Learn phonological process
7. — **Else do**
8. —— Lexicalize exceptions

For example, consider the PL suffix after the first 2 (of 9) inputs listed in Tab. 4.1 have entered

[3]Code is available at https://github.com/cbelth/underlying-forms-SCiL

[4]If two forms are equally frequent, the choice of UR is arbitrary; we used lexicographic ordering to make the ordering complete.

Meaning	UR	Plural Form
PL	/lɑr/	N/A
‘ice’ ‘girl’	/buz/ /kɯz/	**Stem-PL**
‘hand’	/el/	/el-ler/

Table 4.2: When the first three words from Tab. 4.1 enter the lexicon, the stems and plural affix are all stored concretely (left two columns). The plural form of the ‘ice’ and ‘girl’ stems are predictably decomposable into their concrete stems and the PL affix (denoted with the boldface concatenation), so those forms need not be stored in the lexicon. However, with /-lɑr/ as the UR of the plural, the plural form of ‘hand’ cannot be so decomposed, so it is instead lexicalized.

the learner’s lexicon. At this point, the model will be storing the only attested surface form [-lɑr] as the concrete UR /-lɑr/.

When the third word enters the lexicon, our model will lexicalize the form ‘hand-PL’ as /el-ler/, rather than immediately constructing an abstract PL morpheme. This is shown in Tab. 4.2, where each stem and the plural affix have concrete underlying forms, and the plural form of ‘ice’ and ‘girl’ are formed by suffixing the plural to the stem, but the plural form of ‘hand’ is lexicalized.

By the time all 9 words enter the lexicon, however, there will be 4 instances of [-lɑr] and 4 of [-ler], making it no longer sustainable to keep a concrete underlying form. The difference between these two scenarios and, more generally, the decision of when to create an abstract underlying form, is made by the Tolerance Principle (Yang, 2016), as described next.

4.2.4 When is Abstraction Needed?

In order to detect when the amount of surface alternation that prohibits generalization from concrete representations, the model uses the Tolerance Principle, proposed by Yang (2016) and discussed in § 2.5. The Tolerance Principle is a cognitively-grounded tipping point, which hypothesizes that children form productive generalizations when the number of exceptions to a proposed generalization results in a real-time processing cost lower than that without the generalization. The exact derivation of the Tolerance Principle is provided by Yang (2016, ch. 3), but rests critically upon the empirical observation of linguistic sparsity. The Tolerance Principle has had much prior success in computational modeling, lexical, and experimental studies (Schuler et al., 2016; Yang, 2016; Richter, 2018; Koulaguina and Shi, 2019b; Emond and Shi, 2021; Richter, 2021; Belth et al., 2021; Payne, 2022).

Our model’s default treatment of underlying forms as concrete can be stated as a morpheme-specific rule. In the example above, where only the first 2 words of Tab. 4.1 have entered the

Morphemes		Word Forms		
Meaning	UR	PL Form	GEN From	PL, GEN Form
PL	/lɑr/	N/A	N/A	N/A
GEN	/in/	N/A	N/A	N/A
‘ice’	/buz/		??	??
‘girl’	/kɯz/	**ROOT-PL**	??	??
‘stalk’	/sɑp/		??	??
‘hand’	/el/	/el-ler/	??	??
‘word’	/søz/	/søz-ler/	??	??
‘face’	/jyz/	??	/jyz-yn/	??
‘place’	/jer/	??	??	/jer-ler-in/
‘branch’	/dɑl/	??	??	/dɑl-lɑr-ɯn/
‘rope’	/ip/	??	??	/ip-ler-in/

Table 4.3: The left two columns contain morphemes—meaning and form (UR); the right three columns contain word forms. Boldface denotes word forms that can be predictably decomposed into concrete underlying forms, while ‘/-/’ notation denotes word forms that must be lexicalized. The ‘??’ denotes word forms that are unknown. Once all nine words from Tab. 4.1 enter the lexicon, most forms (6 of 9) cannot be predictably decomposed into concrete underlying forms, so the model constructs abstract URs, as described in § 4.2.5.

lexicon, the rule for the PL form would be (72), which predicts that the PL morpheme is realized as [-lɑr].

(72) If PL then [-lɑr]

The Tolerance Principle threshold, which evaluates a linguistic rule (generalization), is repeated from (15) in (73), where n is the number of items the rule applies to and e is the number of exceptions to the rule.

(73)
$$e \leq \frac{n}{\ln n}$$

Thus, our model tracks—for each morpheme—the number of observed words in which the morpheme appears (n) and the number of those where surface alternation leads the morpheme to be realized as something other than its hypothesized concrete form (e).

If the (73) threshold is met, then the UR remains concrete and the word forms where the suffix is realized as something else are lexicalized[5] as exceptions. For example, when the 3rd item in Tab. 4.1 enters the lexicon, the realization of PL as [-ler] violates (72). However, with only three word forms containing PL this one exception can be lexicalized, since $1 \leq 3/\ln 3$.

On the other hand, if the (73) threshold is violated—i.e., $n > \frac{n}{\ln n}$—then the model constructs an abstract underlying form. For example, when the 9th item of Tab. 4.1 enters the lexicon, the realization of PL as [-ler] becomes the 4th of 8 forms in which PL is realized as [-ler] instead of the [-lɑr] predicted by (72). Because $4 > 8/\ln 8$, the model will construct an abstract UR for the PL morpheme.

This is shown in Tab. 4.3, where the plural is realized as [-lɑr] in 3 plural forms and 1 plural, genitive form, but there are 4 forms that must be lexicalized because they instead have the [-ler] form.[6]

Constructing abstract URs introduces discrepancies between URs and SRs for any word forms containing the morpheme, so our model then passes the (UR, SR) pairs implicit in its lexicon[7] to a model that learns phonological alternations to account for the newly-introduced discrepancies. The process of constructing abstract URs is described in § 4.2.5 and the process of learning what conditions the alternations is described in § 4.2.6.

[5]By lexicalization, we mean that the word form is stored in the lexicon verbatim instead of being decomposed into the underlying morphemes. See Tab. 4.2 for an example.

[6]Note that the PL, GEN of 'branch' is lexicalized because the GEN affix is realized in a form other than [in], not because of the PL affix, which is why that form does not get counted as an exception in the Tolerance Principle calculation for the PL affix.

[7]See § 4.2.6 for a description of how the set of (UR, SR) pairs is computed.

4.2.5 Constructing Abstract URs

The model's first step in constructing an abstract UR for a morpheme is to create the set of forms that the morpheme is realized as. For example, the forms of the GEN affix attested in Tab. 4.1 are [-in] / [-ɯn] / [-yn], and of the PL affix are [-lɑr] / [-ler].

Next, the model aligns each of the forms. This is trivial for fixed-length affixes (e.g., the case of the PL affix). If the length of the forms are not all the same, then the model counts the lengths of the morpheme's realizations. For example, the dative affix can be realized as [-ɑ] or [-e], but may contain an affix-initial [j] when attaching to a morpheme that ends in a vowel. The model thus counts the number of words in which [-ɑ] or [-e] (length 1) is the realization, and the number in which [-jɑ] or [-je] is the realization (length 2), and chooses the most frequent length as the length of the UR. If a shorter length is chosen, the extra segment(s) are treated as epenthesized; if the longer is chosen, they are treated as deleted. For simplicity, we assume that these segments epenthesize or delete on the left, which is a simplification. This process is not guaranteed to generalize to other languages, so future work will develop a more robust alignment process by more tightly combining the problems of abstract UR construction and rule construction.

Once the forms are aligned, the UR is constructed one segment at a time. Each segment is set to match in features where all realizations of the affix match; features that alternate across forms are unspecified underlyingly. For example, [-lɑr] / [-ler] will lead to /-lAr/, where A is the low, unround vowel with backness unspecified, because both forms agree in the initial and final segments, but the vowel alternates on backness. Similarly, [-in] / [-ɯn] / [-yn] will result in /-Hn/, where H is the high vowel with backness and roundness unspecified, since [i] and [y] differ in backness from [ɯ] while [i] and [ɯ] differ from [y] in roundness.

4.2.6 Learning Alternations

When the number of words where the morpheme's surface alternation requires the word be lexicalized becomes too great, the model constructs an abstract UR for the morpheme. This abstract UR introduces a discrepancy between the abstract UR and its surface realization. The model thus constructs a set of (UR, SR) pairs from the lexicon, which it passes to a model that learns a phonological process to derive the various surface forms.

For example, when the 9th item from Tab. 4.1 causes /lɑr/ to no-longer be sustainable as the PL affix UR, the lexicon is as described in Tab. 4.3. The surface form for the PL forms of the roots 'ice', 'girl', and 'stalk' are computed by concatenating /lɑr/ to the stem (i.e., Stem-PL), and the remaining six known surface forms, which were lexicalized, are extracted directly from the lexicon. Since the PL is being collapsed into /lAr/, each word's UR is computed by replacing the surface realization of the PL affix with this new UR. Thus, the (UR, SR) pairs at this point would be {(/buzlAr/, [buzlɑr]),

(/kɯzlAr/, [kɯzlɑr]), (/sɑplAr/, [sɑplɑr]), (/ellAr/, [eller]), (/søzlAr/, [søzler]), (/jyzyn/, [jyzyn]), (/jerlArin/, [jerlerin]), (/dɑllArɯn/, [dɑllɑrɯn]), (/iplArin/, [iplerin])}.

Learning phonological processes from UR-SR pairs is an active area of study, and many models have been proposed (see Jarosz 2019 for an overview). In this chapter we chose D2L (§ 3), which is a cognitively-grounded model that provides a unified ability to learn local and non-local alternations, which is important, given Turkish's non-local vowel harmony combined with local processes like voicing assimilation (see § 4.3.1). See chapter § 3 for a description of the D2L model.

4.3 Turkish Case Study

This section provides a case study of our proposed model on the highly agglutinative language, Turkish. In § 4.3.1 we describe some relevant details of Turkish. We then describe the setup of our evaluation in § 4.3.2. Finally, we present qualitative results in § 4.3.3 and quantitative results in § 4.3.4.

4.3.1 Turkish

Turkish phonology receives attention often because of its apparently complex vowel harmony system. It exhibits both primary front/back harmony and secondary rounding harmony, which is parasitic on height: only [+high] vowels harmonize for roundness. Moreover, Turkish has a number of exceptional suffixes whose vowels do not participate in harmony, and even half-harmonizing suffixes, which have multiple vowels, some of which harmonize and some of which do not. These harmony processes occur alongside other processes, such as local voicing assimilation. The Turkish vowel inventory is shown in (74).

(74)

	front		back	
	unround	round	unround	round
high	i	y	ɯ	u
low	e	ø	ɑ	o

The primary harmony process is observed in affix vowels that alternate between [+back] when the preceding vowel is [+back] and [−back] when it is [−back], as in (75) (examples from Nevins 2010, p. 28; Kabak 2011, p. 3).

(75)

[dɑ-lɑr-ɯn]	branch-PL-GEN
[jer-ler-in]	place-PL-GEN
[ip-ler-in]	rope-PL-GEN

The secondary rounding harmony involves suffixal [+high] vowels matching in roundness to the vowel to the left, as in (76) (examples from Nevins 2010, p. 29; Kabak 2011, p. 3). This harmony occurs regardless of whether the vowel to the left is [+high] (76a) or [−high] (76b).

(76)	a.	[ip-in]	rope-GEN
		[jyz-yn]	face-GEN
		[kɯz-ɯn]	girl-GEN
		[buz-un]	ice-GEN
	b.	[el-in]	hand-GEN
		[søz-yn]	word-GEN
		[sɑp-ɯn]	stalk-GEN
		[jol-un]	road-GEN

Some suffixes have vowels that do not participate in harmony.[8] For example, the suffix [-ki] can attach to a temporal or spatial nominal root to yield adjectival forms as in (77), where the suffix surfaces with the vowel [i] regardless of the final vowel of the stem (examples from Oflazer 1994, p. 144). The PL suffix, which alternated in (75), here harmonizes with the [i] vowel (77b), thus surfacing as [e].

(77)	a.	[øndʒe-ki]	'(the one) before'
		[jɑrɯn-ki]	'(the one) tomorrow'
	b.	[øndʒe-ki-ler]	'(the ones) before'
		[jɑrɯn-ki-ler]	'(the ones) tomorrow'

The situation gets more complex, as some suffixes are *half harmonizing*, meaning they have two vowels with one harmonizing and one not.[9] An example is shown in (78a), where the first vowel of the abilitative (ABIL) suffix harmonizes with the vowel to the left, but the second vowel is always [−back] [i] even when the first vowel is [+back] (Kornfilt, 2013). The aorist (AOR) suffix vowel then harmonizes with the abilitative's non-harmonizing second vowel [i] in (78a). The example (78b) demonstrates that the AOR suffixal vowel surfacing as [i] in (78a) is indeed due to harmony, as it harmonizes in (78b) with [o].

(78)	a.	[jɑz-ɑbil-ir]	'write'-ABIL-AOR
		[jyz-ebil-ir]	'swim'-ABIL-AOR

[8]One reviewer of our Belth (2023c) paper pointed out that such vowels can be called *opaque*. We chose to eschew the term in favor of describing the behavior of such vowels directly, to increase readability for the uninitiated, and to leave the term *opaque* for vowels in harmony systems in which they behave opaquely throughout the system.

[9]The term *half harmonizing* is from Nevins (2010), but one reviewer of our Belth (2023c) paper pointed out that, in principle, other fractions of vowels (1 of 3) could harmonize.

b. [ol-ur] ‘become’-AOR

Vowel harmony often goes in hand with other phonological processes, such as voicing assimilation. This can be seen, for example, in the locative (LOC) suffix, which exhibits vowel harmony, but begins with an alveolar stop, which assimilates in voicing to the segment to its left, as in (79) (examples from Dobrovolsky 1982; Çöltekin 2010; Kornfilt 2013).

(79) [byro-dɑ] ‘office’-LOC
[ev-de] ‘house’-LOC
[ʤep-te] ‘pocket’-LOC

In the remaining subsections, we demonstrate how our proposed model elegantly accounts for these complexities in Turkish (§ 4.3.3), and how this allows for novel surface forms to be accurately predicted (§ 4.3.4). First, though, we introduce the setup and data we used for our experiments (§ 4.3.2).

4.3.2 Setup and Data

To simulate learning in Turkish, we constructed two Turkish datasets consisting of frequency-annotated and morphologically-analyzed surface forms (see below). To simulate one learning trajectory, we sampled words with replacement from the corpus, weighted by frequency. Each time a new word form is sampled, the learner adds it to its lexicon. We then investigate the underlying forms of each morpheme, seeing which are concrete and which are abstract (§ 4.3.3). We then evaluate how accurately the model, combined with a model for learning alternation rules, allows novel surface forms to be predicted (§ 4.3.4).

We constructed two datasets, called MorphoChallenge and CHILDES. The first used data from MorphoChallenge (Kurimo et al., 2010), which contains a large Turkish corpus annotated with word frequencies. To generate morphological analyses of words, we used Çöltekin (2010, 2014)’s finite state morphological analyzer, which is designed for Turkish. This is similar to the process used in the MorphoChallenge, but is publicly available.[10] We dropped any word in MorphoChallenge that had fewer than 25 occurrences or for which the morphological analyzer failed to provide an analysis. We also removed forms with affixes that are analyzed by Çöltekin (2010, 2014) as having multiple underlying forms. For example, the highly irregular aorist suffix is sometimes described as having four underlying forms: /-Ar/, /-Hr, /-z/, /-null/. Future work will consider scenarios where multiple URs are necessary. This resulted in 22,315 frequency-annotated and morphologically-analyzed surface forms, which we transcribed into IPA.

[10] https://github.com/coltekin/TRmorph

The second dataset is derived from the child-directed speech in the Aksu (Slobin, 1982) and Altinkamis corpuses of the CHILDES database MacWhinney (2000). We computed the frequency of each word in the corpuses and used the same process as above to morphologically analyze each word. This dataset is much smaller, so we did not exclude words with low corpus counts from this dataset. The resulted in 1,727 frequency-annotated and morphologically-analyzed surface forms, transcribed into IPA.

Note that some Turkish suffixes exhibit deletion/epenthesis to avoid CC or VV clusters. These additional processes are at present ignored, because the implementation provided by § 3 was designed for harmony and disharmony. Future work will extend the implementation to epenthesis and deletion by incorporating PLP, which handles such processes.

4.3.3 Suffixes: Abstract and Concrete

Remarkably, the apparent complexity of Turkish vowel harmony, discussed in § 4.3.1, vanishes when we investigate the output of our model.[11] As before, we will let A denote the Turkish low, unround vowel with backness unspecified (extensionally, {e, ɑ}) and H be the Turkish high vowel with both backness and height unspecified (extensionally, {i, y, ɯ, u}). Moreover, we will use D to denote the alveolar stop with voicing unspecified (extensionally, {d, t}).

We will walk through the complexities exemplified by (75)-(79) one-by-one. First, the PL suffix in (75), which has a low unrounded vowel, participates in front/back harmony, but not rounding harmony because it is not a [+high] vowel. Our model constructed the underlying form /-lAr/ for this suffix, capturing the fact that it only harmonizes for backness.

The GEN suffix in (75)-(76) has a [+high] vowel and participates in both primary and secondary harmony. Our model constructed the underlying form /-Hn/ for this suffix, which captures the surface alternation of this morpheme.

Next, the [-ki] suffix in (77) does not participate in harmony, and our model consistently represents it with a concrete form /-ki/.

For the abilitative suffix in (78), our model abstracts the first, harmonizing vowel, but keeps the second, non-harmonizing vowel concrete -/Abil/.

Lastly, the UR for the locative suffix in (79) is constructed with both segments abstract /-DA/, capturing both the voicing assimilation of the initial alveolar stop and the vowel harmony of the second segment.

These underlying forms allow D2L to learn two rules, which allow for the accurate prediction of novel surface forms. On the resulting (UR, SR) pairs, D2L learns a vowel harmony rule, which targets both /A/ and /H/ vowels, and enforces harmony with respect to their unspecified values:

[11]This analysis is performed on a random, frequency-weighted 80% sample of the MorphoChallenge dataset.

Affix	UR	Abstract
PL	/-lAr/	Y
P3S	/-H/	Y
P1S	/-m/	N
GEN	/-Hn/	Y
DAT	/-A/	Y
ACC	/-H/	Y
LOC	/-DA/	Y
VN:INF	/-mA/	Y
lH	/-lɯ/	N
P2S	/-n/	N

Table 4.4: Top 10 most frequent affixes in a random, frequency-weighted sample of 1K words from the CHILDES dataset, and the URs that our model learned. See http://coltekin.net/cagri/trmorph/trmorph-manual.pdf for a description of affix names.

[back] for /A/ and both [back] and [round] for /H/ (80a). The model automatically constructs a vowel tier and enforces harmony locally over that tier (see § 3 for details). D2L also learns a local voice assimilation rule, which causes /D/ to take its [voi] value from the segment to its left (80b).[12]

(80) a. AGREE([?back], {back, round})/[−cons] __ ∘ proj(·, [−cons])

b. AGREE([?voi], {voi})/[∗] __

It is worth noting that others—in particular Nevins (2010)—have similarly argued that Turkish vowel harmony can be elegantly accounted for with an underspecification approach. Our model builds on Nevins (2010)'s observations by providing an explicit computational model that constructs underlying forms, which turn out to be consistent with this analysis.

As a further analysis, we show the 10 most frequent affixes in a 1K word sample of the CHILDES corpus in Tab. 4.4, along with the UR that our model constructed for each. Of the 10 affixes, 7 have been collapsed into abstract forms. However, there are 3 forms (P1S, lH, P2S) that were quite frequent, but are still able to be stored concretely. The P1S and P2S affixes do not have alternating segments in Turkish, so it is expected that these would be concrete. The "lH" affix, as captured by its name, can surface with any high vowel. However, in the training data, the [-lɯ] form occurs 25 out of 32 times, so the 7 words where it surfaces as something else are lexicalized ($7 <= 32/\ln 32$).

[12]The notation [*] denotes any segment.

4.3.4 Quantitative Evaluation

We also evaluated how the model enables generalization, when paired with a model for learning phonological alternations. We used our model in tandem with D2L to learn to map a stem and morphological analysis of a surface form to an actual surface form. For example, given the stem [dɑl] and morphological analysis Stem-PL-GEN, our model's underlying forms for -PL and -GEN are concatenated to the stem to form a UR, to which the generalizations learned by D2L can then be applied to predict a surface form, such as [dɑllɑrɯn].

4.3.4.1 Setup

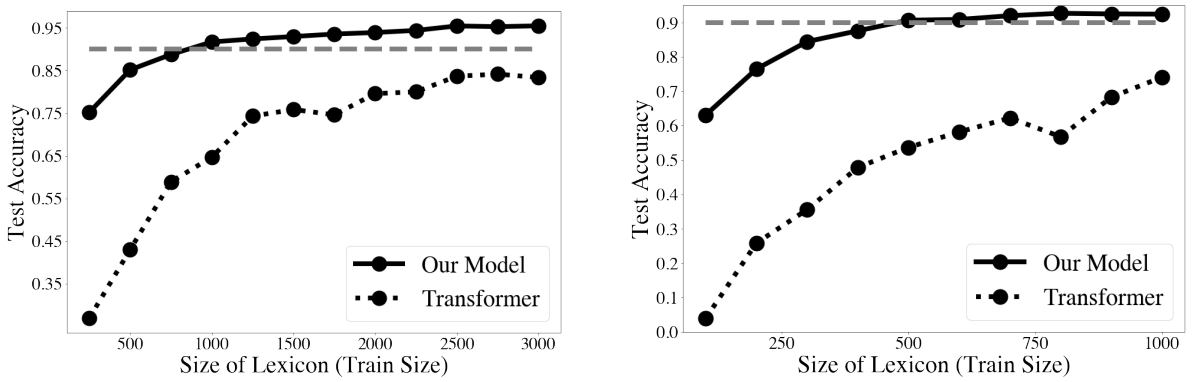
We ran the model on both datasets, simulating incremental learning by sampling words with replacement and weighted by frequency, and adding them to the lexicon when sampled. As this process incrementally adds words to the lexicon, our model operates as described in (71). In 250-word increments (i.e., every time the lexicon grows by 250 unique words) for MorphoChallenge and 100-word increments for CHILDES, we evaluated the model by using the rules learned by Belth (2023b)'s model—on our learned underlying forms—to predict the surface form of all the words not in the lexicon. We carried out 5 simulations on each dataset, using different random seeds for sampling on each.

As a comparison, we used a transformer-based (Vaswani et al., 2017) seq-to-seq model. At each training increment (250 words for MorphoChallenge and 100 for CHILDES), we created a random 80/20 train/dev split of the training data to tune hyperparameters. We search over the hyperparameters shown in (81) and choose the combination with the highest accuracy on the dev set. We then trained a model with the best hyperparameters on the entire training set (i.e. re-merging the 80/20 split).

(81) learning rate $\in [0.0001, 0.01]$
number of epochs $\in \{10, 11, 12, ..., 29, 30\}$
embedding dimension $\in \{16, 32, 64, 128, 256, 512\}$
hidden dimension $\in \{16, 32, 64, 128, 256, 512\}$
number of attention heads $\in \{1, 2, 4, 8\}$
number of encoder layers $\in \{1, 2, 3, 4\}$
number of decoder layers $\in \{1, 2, 3, 4\}$

4.3.4.2 Results

The results are shown in Fig. 4.1, where the x-axis shows the incremental growth of the learner's lexicon (i.e., the training size), and the y-axis shows the accuracy at predicting novel surface forms



(a) MorphChallenge

(b) CHILDES

Figure 4.1: Our proposed model's accuracy (averaged over 5 simulations) at predicting novel surface forms. The x-axis shows the growth of the learner's lexicon (i.e., the training size). The gray, dashed line marks the 0.9 accuracy point.

at that point during training. The accuracy is computed over all surface forms not currently in the training data. Each subfigure is for one of the two datasets. The MorphoChallenge results (Fig. 4.1a) are reported up to a size of 3K words, so the test results are on 10s of thousands of novel words.

Our model's accuracy is higher than that of the transformer model at all training sizes. The dip in transformer accuracy at 800 training size in CHILDES is likely due to high variance across the 5 runs, and the curve would likely smooth with more simulations.

Our model's performance appears to be consistent with acquisition studies. Altan (2009) found that Turkish-speaking children as young as 2;0 extend vowel harmony to nonce words. Studies across languages reveal that a child's vocabulary is quite modest at this age, with an upper bound around 1K words (Fenson et al., 1994; Hart and Risley, 1995; Szagun et al., 2006; Bornstein et al., 2004). The model's performance on both datasets is above 90% accuracy (gray, dashed line) when its vocabulary contains 1K words.

4.3.4.3 Error Analysis

Of the errors, around 52% result from the model having a concrete form of an affix, which it then errantly predicts for a novel word that exhibits alternation in that affix. For example, there are insufficient forms in the training data to make /ɯp/ as the concrete CV:IP affix prohibitive (e = 5 <= n = 13/ln 13), even though vowel harmony leads it to sometimes surfaces with other high vowels. As a result, novel words like [gel-ip], which take the [ip] form of the affix are mispredicted.

About 47% of the errors are the result of vowel harmony or consonant assimilation being predicted for a novel form that exceptionally does not involve harmony. For example, the word [saat-ler] 'watch-PL' is predicted by our model to be [saat-lar] because the UR for the plural suffix is

/lAr/, as it systematically harmonizes. According to a Wiktionary search,[13] the root [saat] is of Arabic origins. Because Arabic has a different vowel system, vowels in Arabic loan words may conform to the Turkish vowel system when entering Turkish, and thus sometimes behave oddly. Indeed, Altan (2009) observed that children may overextend vowel harmony to such words.

The remaining 1% of errors result from very low frequency affixes which are simply unattested in the training data.

4.4 Dutch Case Study

In this section, we provide a case study of our proposed model on voicing alternations in Dutch noun paradigms. In § 4.4.1 we describe the relevant details of Dutch. In § 4.4.2 we describe our experimental setup. Then, in § 4.4.3 we discuss how the model's results may account for experimental results discussed in § 4.4.1, and demonstrate that the model's output allows accurate prediction of novel forms in § 4.4.4.

4.4.1 Dutch

Dutch, like many other languages (e.g., German and Polish), exhibits a phonotactic restriction against syllable-final voiced obstruents. Beyond the distribution of voiced obstruents, a primary indication of this restriction comes from certain informative noun paradigms. Many Dutch plural nouns are formed by suffixing [-ən]. When this plural suffix attaches to a stem ending in an obstruent, the obstruent is syllabified as the onset of the syllable containing the plural suffix syllable. In some noun paradigms—such as (82a; data from Zamuner et al. 2012, p. 482)—the stem-final obstruent is voiced in the plural form, but unvoiced in the singular, where it occurs in syllable-final position. In other paradigms, the obstruent is voiceless throughout the paradigm—e.g., (82b).

(82) a. [bɛt] 'bed' [bɛdən] 'bed'-PL
b. [pɛt] 'cap' [pɛtən] 'cap'-PL

Because not all nouns with stem-final obstruents alternate, researchers have taken interest in whether Dutch learners are aware of which nouns alternate and which do not (i.e., are they sensitive to the alternation for nouns that they know?), and, if so, whether learners generalize productively so as to expect stem-final voiced obstruents in novel plural nouns to be voiceless in the singular.

Dutch nouns also frequently occur in the diminutive (both singular and plural), with especially common occurrence in child-directed speech. In fact, some plural diminutive forms of a noun are more frequent than their non-diminutive counterparts. For example, Kerkhoff (2007, p. 113) found

[13] https://en.wiktionary.org/wiki/saat#Turkish

that *eendjes* 'ducklings' occurs more times (eight) in her corpus of child-directed speech than does the *ducks* (three). This is an important fact[14] because nouns with voiced stem-final segments in the (non-diminutive) plural are unvoiced in both the singular and plural diminutives, as in the (non-diminutive) singular. For example, in (83) the stem-final obstruent of *paard* 'horse' surfaces as [t] in all forms except for the non-diminutive plural (83b).

(83) a. [part] 'horse'
b. [pardən] 'horses'
c. [partjə] 'horsie'
d. [partjəs] 'horsies'

If the child learns (83c) 'horsie' and (83d) 'horsies' before learning (83b) 'horses'—a plausible scenario given the prevelance of dimnuitves in child-directed speech—then the child receives no evidence of an alternation. This highlights the importance of considering the details of the child's input in understanding what representations and generalizations they may construct.

Many other languages exhibit syllable-final devoicing that manifests in voicing alternations. For example, German exhibits a descriptively similar alternation, in which devoicing of obstruents in final position appears in the form of voicing alternations like [hʊnt] 'dog' ~ [hʊndə] 'dogs.' Because of their descriptive similarity, these cross-linguistic patterns and processes are usually treated as a group. However, what is cross-linguistically descriptively similar may not have the same cognitive status across speakers of the languages.

In an age-controlled experiment, Buckler and Fikkert (2016) found German 3-year-olds show more advanced knowledge of which nouns in their mental lexicon alternate in comparison to 3-year-old Dutch children. The authors argue that this is likely attributable to language-specific factors, including the fact that German uses voicing contrastively over a broader range of obstruents than Dutch does and that Dutch exhibits both progressive and regressive voicing assimilation while German only commonly exhibits progressive voicing assimilation. These cross-linguistic differences may make the voicing alternating in noun paradigms harder to acquire for Dutch learners than German learners. Perhaps more significantly, Buckler and Fikkert (2016) performed a corpus study of German and Dutch nouns in CHILDES (MacWhinney, 2000), and found that the analyzed German child-directed speech contained a much greater number of alternating noun paradigms than that of Dutch, whether counted in type frequency (72 in German vs. 22 in Dutch) or token frequency (1572 in German vs. 158 in Dutch).

Several other studies have investigated whether knowledge of the Dutch voicing alternation is productive for children, and found evidence that it is not. Zamuner et al. (2012) performed a reverse wug test with 2.5-year-old and 3.5-year-old children. A reverse wug test presents the plural form

[14]One that is, oddly, not often discussed in experimental investigations of the Dutch voicing alternation.

of a nonce noun and asks participants to produce the singular, which flips the classical setting originating from Berko (1958), in which the participants are asked to produce the plural form from the singular nonce. For example, the children were presented with nonce noun plurals where the stem-final obstruent for some was voiced [d], as in (84a), and for others was voiceless [t], as in (84b).

(84) a. [slɑdən]
[kɛdən]

b. [klatən]
[jɪtən]

The nonce-paradigms with the [d] plural forms alternate, since the [d] will occur in final position in the singular; the paradigms with [t] plurals do not alternate.

In general, children at both ages were not very good at producing a singular form at all, perhaps showing the difficulty of the task for a young child. However, when children did produce a singular form, they showed significantly better performance at both ages when the stem-final obstruent was voiceless—i.e. the nonce noun paradigm did not alternate—than when it was voiced.

Another study by Zamuner et al. (2006) corroborates these results. The structure of Zamuner et al. (2006)'s study was very similar to that of Zamuner et al. (2012), but it tested children's comprehension rather than production. Again, 2.5 year old and 3.5 year old children were tested. The children were presented the same stimuli as in the production study, but instead of being asked to produce the singular form, an experimenter read the singular form and asked the children to point to the correct image. There were three image choices: the plural image of the nonce, the singular image of the nonce (the correct choice), and a distractor (a plural image of a different nonce). For example, a child would be shown a plural picture and told that it represents multiple [slɑdən], and then asked to point to the picture of a [slɑt].

Children showed better performance at this comprehension task than production-based tasks. Moreover, children at both ages were much more accurate at identifying the correct singular image when the nonce-paradigm did not alternate than when it did. The authors also flipped the experiment, performing a standard singular-to-plural (comprehension) wug test. Again, the children were significantly better at identifying the correct plural for non-alternating paradigms.

These results are perhaps not surprising. Buckler and Fikkert (2016) found Dutch 3-year-olds to show little knowledge of which nouns in their mental lexicon alternate; lacking such robust knowledge, there would be little information for children to have used to construct a productive generalization for the voicing alternation.

Kerkhoff (2007, ch. 6) found that adult speakers of Dutch were able to produce singulars for nonce plurals from both alternating and non-alternating paradigms. This may constitute evidence

that the voicing alternation eventually emerges as productive. However, such productivity may be aided by orthography: Kerkhoff (2007) observes that, e.g., the orthographic rendering of (82a) *bed* ~ *bedden* makes the status of the final [t] in [bɛt] as a neutralized /d/ transparent, and children continue to show little evidence of productively generalizing the alternation even at age 6.

In the remainder of this chapter, we will demonstrate how our model gives a possible learning-based explanation for these results. When run on a realistic child-directed speech corpus, the prevalence of voicing alternations in noun paradigms is rare enough that alternating forms can be lexicalized as the exceptions to productive morphological rules. Consequently, no phonological process is constructed.

Before proceeding, it is important to address one remaining fact about Dutch nouns. Some nouns are formed by suffixing [-s], not [-ən], to the stem; (85) gives examples.

(85) [voːɣəl] 'bird' [voːɣəls] 'birds'
[vɪŋər] 'finger' [vɪŋərs] 'fingers'

The [-s] plural is usually used after syllables that do not have main stress (Booij, 1999, p. 82). Because [ə] does not bear stress in Dutch (Booij, 1999, p. 5), a very strong indicator of the [-s] plural suffix is a stem ending in a syllable with a [ə] nucleus. The [-s] suffix is less frequent than the [-ən] suffix, does not often attach to a stem with a final obstruent, and does not lead to a stem-final consonant being re-syllabified. Thus, the [-s] suffix does not directly contribute to the Dutch voicing alternation, and nouns taking it are usually ignored in corpus and experimental studies. However, a complete story of Dutch nouns must account for these, and our own dataset (see below) retains them.

4.4.2 Setup and Data

We extracted nouns from the child-directed speech in the Van Kampen (Van Kampen, 1994, 2009) corpus of the CHILDES (MacWhinney, 2000) database. We used all sessions with children up to age 3.5 years, following Buckler and Fikkert (2016)'s corpus study, and the age of children in Zamuner et al. (2006, 2012)'s experiments. We used the TreeTagger[15] part-of-speech tagger (Schmid, 1999, 2013) to extract nouns from the corpus. The tagger also specifies whether the noun is singular or plural. We tagged diminutive nouns post-hoc based on the *-je* suffix (the TreeTagger provides lemmas for tagged words). To get IPA transcriptions, we took the intersection of the resulting set of nouns with the Dutch part of *wikipron* (Lee et al., 2020). Finally, we computed the frequency of each word in the corpus, and dropped words with only a single occurrence. This resulted in a set of 887 nouns—606 singular, 107 plural, 124 singular diminutive, and 50 plural diminutive.

[15] https://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/

Meaning	UR	Abstract
PL	/-ən/	N
DIM	/-jə/	N
'horse'	/part/	N
'book'	/buk/	N

Table 4.5: The underlying forms of the plural and diminutive suffixes, as well as example roots 'horse' (which alternates) and 'book' (which does not alternate) after running on the model on the 887 Dutch nouns from the CHILDES dataset.

4.4.3 Generalization Without Abstraction

We ran our model on the 887 nouns in our dataset. The underlying forms of the plural (PL) and diminutive (DIM) affixes, as well as the root nouns 'horse' and 'book' are shown in Tab. 4.5. All four morphemes are concrete. Even though the noun 'horse' occurs in singular, plural, singular diminutive, and plural diminutive forms, the root occurs as [part], with voiceless [t], in all but the plural [pardən] (see. 83). Since $1 \leq 4/\ln 4$, the plural form is lexicalized. The root contains voiceless /t/ because 3 of 4 forms realize the obstruent as that. Similarly, the UR for 'book' is [buk] because this is the realization of the root in all four forms.

In short, then, surface alternation is rare enough in our sample of Dutch child-directed speech that concrete forms are sustainable for root morphemes, even if the final segment is an alternating obstruent. It is possible, however, that the number of alternating noun paradigms could grow large enough that it prohibits the formation of morphological (re-)inflection rules. For example, suppose a learner is trying to learn a generalization from SGs to PLs and posits the rule (86), stating that the plural form of a noun can be derived by suffixing [-ən] to the root form.

(86) PL = ROOT-[ən]

Alternating noun paradigms will appear to be exceptions to this generalization, since suffixing [ən] to a root form that ends in a voiceless obstruent will fail to yield the appropriate plural form in such cases (e.g., producing *[partən] instead of [pardən]). This is not a feature of *particular* morphemes, and hence does not lead our model to posit abstract underlying forms. It is, however, a possible second stage at which abstraction could become necessary.

To determine whether this is the case for Dutch, we ran the *Abduction of Tolerable Productivity* (ATP) model from Belth et al. (2021), to learn morphological inflection rules from the underlying forms our model constructs. The model learns to map from lemmas (root forms) and features (morphosyntactic and/or phonological) to an inflected form.

If the model fails to construct a rule for deriving the plural form from the root, this is evidence that abstraction is necessary. Future work will combine these into a single model.

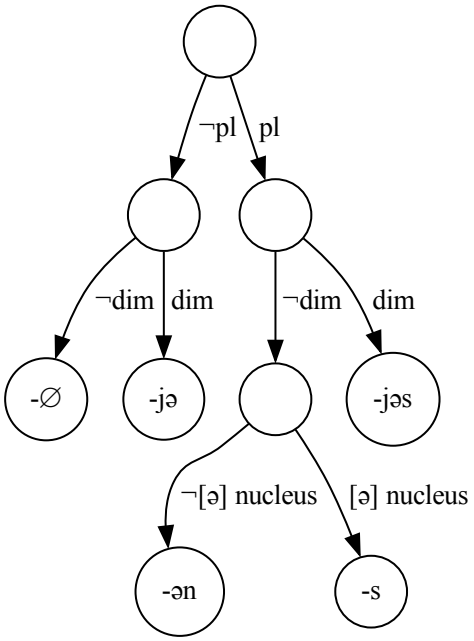


Figure 4.2: Morphological Inflection Rules

ATP's output is shown as a decision tree in Fig. 4.2. The left-most leaf shows that (non-diminutive) singulars add no suffix (-∅ denotes the null suffix) to the lemma. The [-jə] and [-jəs] leaves show that singular and plural diminutives are formed by suffixing [-jə] and [-jəs] to the lemma, respectively. The remaining two leaves show how plurals are formed: [-s] if the lemma ends in a syllable with a [ə] nucleus, and [-ən] otherwise.

In (87), we show the Tolerance Principle counts for each of ATP's rules. The rules are shown as an ordered list from most specific (deepest in tree) to least specific.

(87) a. PL ∧ ¬DIM ∧¬[ə] nucleus → [-ən] ($n = 90, e = 10 \leq 90/\ln 90 = 20.0$)

b. PL ∧ ¬DIM → [-s] ($n = 17, e = 2 \leq 16/\ln 16 = 6.0$)

c. PL → [-jəs] ($n = 50, e = 0 \leq 50/\ln 50 = 12.8$)

d. ¬DIM → -∅ ($n = 603, e = 0 \leq 603/\ln 603 = 94.2$)

e. Elsewhere → [-jə] ($n = 124, e = 6 \leq 124/\ln 124 = 25.7$)

Since alternations only occur for paradigms where a stem-final obstruent is re-syllabified in the onset of the [-ən] PL suffix, the alternating nouns fall in the third-from-left branch of the tree, whose counts are shown in (87a). Since $10 \leq 90/\ln 90$, the rule is productive despite the alternation and the

Table 4.6: Both humans and our model perform much better at identifying the singular of a nonce plural when the noun does not alternate than when it does. In comparison, a trigram language model does not reflect this asymmetry.

	Non-Alternating	Alternating
Humans	0.61	0.48
Our Model	1.00	0.33
Trigram	1.00	1.00

alternating nouns can be lexicalized. Moreover, of the 10 exceptions, only 8 are due to the voicing alternation.[16] The parent node of the plural suffixes (87a)-(87b) had $n = 90 + 17 = 107$ nouns, of which 16 take the [-s] suffix and only 8 alternate. Thus, subdividing to isolate the 16 nouns that take [-s] instead of subdividing to isolate the 8 alternating nouns that take [-ən] is expected. Note that if the underlying forms of alternating noun stems contained voiced obstruents instead of voiceless obstruents, the exceptions would instead have gone to the left-most branch (87d) (since devoicing would be necessary to produce the correct singular form), but $8 \leq 94.2$, so the conclusion would hold even more strongly.

It appears, then, that Dutch-learning children may lack productive knowledge of voicing alternations because such knowledge is not important to their developing linguistic system. Our model demonstrates that highly productive morphological generalizations are possible without knowledge of voicing alternations, and perhaps child learners are content with these imperfect but fully sufficient generalizations.

To make this point more clear, consider our model's behavior on the nonce nouns from Zamuner et al. (2006, 2012). We consider the comprehension reverse wug test, in which children were presented with a nonce plural followed by its singular form and were then asked to identify the picture corresponding to the singular. Since our model did not need to learn a voicing alternation, when we present the plurals to our model, the predicted singular form is simply the plural with the [-ən] suffix removed. For example, our model predicts [slɑd], [kɛd], [klat], and [jɪt] as the singular for the alternating nonce words in (84), repeated in (88).

(88) a. [slɑdən]
[kɛdən]

b. [klatən]
[jɪtən]

The predictions will match the singular spoken by the experimenter, which always ends in [t],

[16]The others are due to one stem that takes [-s] and one stem that has an irregular vowel change in the plural.

for non-alternating forms (88b) but not for the alternating forms (88a). When the model's prediction does not match the singular, we suppose that the image is chosen at random (from three choices). The resulting predictions are shown in Tab. 4.6, where our model predicts the expected form for non-alternating nouns but performs at-chance for alternating nouns. For reference, we report the average performance of the children in Zamuner et al. (2006)'s study.[17] Clearly the children performed much worse overall than the model, perhaps indicating the overall difficulty of the task for children. However, the trend is correct. Contrast this with a trigram language model fit to the surface forms of the 887 Dutch nouns. Because voiced obstruents never occur in syllable final position, the [t] singular form spoken by the experimenter always has higher probability than the same noun with a final voiced [d]. This highlights how knowledge of the phonotactic restriction against voiced obstruents in final position does not automatically lead learners to create a productive voicing alternation generalization.

4.4.4 Quantitative Evaluation

4.4.4.1 Setup

As for Turkish, we evaluated how the model enables morphophonological generalization. However, as discussed above (§ 4.4.3), the voicing alternation is not pervasive enough to require abstract underlying representations. This was established by running our model on the entire dataset of 887 nouns. It is important to confirm that the these observations are true throughout (simulated) development of the mental lexicon. Secondly, it is important to know that when the calculation of sufficiently accurate generalization from concrete representations (via the Tolerance Principle) does not require abstraction, that morphophonological generalizations constructed over concrete representations still achieve high accuracy generalizing to new words.

Thus, we ran our model on the Dutch dataset, simulating incremental learning. As for Turkish, each simulation sampled words with replacement and weighted by frequency, and added them to the lexicon when sampled. In 100-word increments up to a vocabulary size of 700 nouns, we confirmed that all underlying forms were still concrete and that productive morphological inflection rules could still be extracted via ATP. At each of these increments, we also evaluated the accuracy of the resulting morphological inflection rules over 187 test words not seen during training. We performed 30 simulations using different random seeds.

We compared to a transformer-based (Vaswani et al., 2017) seq-to-seq model. At each training size, we created a random 80/20 train/dev split of the training data and tuned hyperparameters—the

[17] Zamuner et al. (2006)'s results were broken down by age, but no significant effect was found for age, so we aggregated the results into a single number for each type of noun by taking the weighted average of the two age groups' results.

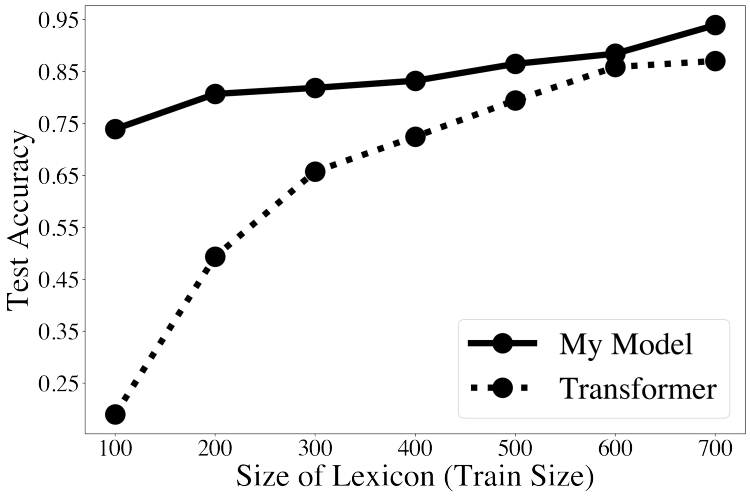


Figure 4.3: Accuracy generalizing to held-out test words.

same as Turkish (81)—by choosing the combination with the highest accuracy on the dev set, and then training a model with the best hyperparameters on the entire training set (i.e. re-merging the 80/20 split).

4.4.4.2 Results

Across the 30 simulations, the voicing alternation was never pervasive enough to require the model to construct abstract underlying forms nor to prevent ATP from learning productive morphological inflection rules. The accuracy generalizing to held-out test words, shown in Fig. 4.3, approaches 0.95 by the time the vocabulary contains 700 nouns. The performance surpasses that of the transformer seq-to-seq model at all training sizes.

Despite not creating abstract underlying forms for alternating paradigms—and hence not learning a productive voicing alternation—the high overall accuracy demonstrates that accurate generalizations are nevertheless possible.

4.5 Prior Work

Tesar (2013) and Hua et al. (2020) focus on theoretical analyses of the nature of the problem of learning URs. O'Hara (2017); Rasin et al. (2018); Ellis et al. (2022) proposed computational models, but evaluate on small, phonology-textbook-like data, not large, natural-language corpora.

Cotterell et al. (2015) also predominately models textbook-like problems, but presents some limited analysis on more realistic corpora. However, these corpora only involve very simple morphological paradigms involving a single suffix, and present to the model a fairly curated subset of

the corpus that isolates the relevant morphophonological process.

Richter (2021) studies the question of when allophonic surface segments are collapsed into an abstract underlying segment, focusing on the English flap [ɾ] allophone of /T/. While Richter (2021) focuses on allophones, our proposed model is inspired by it and can be viewed as extending the same principles to morphophonological alternations.

Of these prior models, we were only able to get access to code for Cotterell et al. (2015) and Rasin et al. (2018), which we were unable to get to run on our large datasets. In future versions of this work, we intend to implement some of these existing models in order to compare their performance and behavior to that of our proposed model.

4.6 Conclusion

This chapter began an algorithmic, learning-based account of underlying forms, taking the highly agglutinating language of Turkish and an apparent lack of productivity in Dutch as two case studies. The proposed model starts with concrete underlying representations and constructs abstract URs only in cases where doing so is needed to form generalizations that deal with the sparsity of morphological forms in the learner's input.

The model constructs abstract underlying forms when they are critical for generalization, but allows for concrete forms when abstraction is unnecessary. This flexibility is at the core of the model's success, as evidenced by the constructed underlying representations of Turkish suffixes in § 4.3.3 and alternating Dutch noun roots in § 4.4.3. For example, the half-harmonizing suffixes consist of concrete segments except for the single, harmonizing vowel. Similarly, exceptional, non-harmonizing suffixes remain fully concrete.

When combined with a model for learning local and non-local alternations, the proposed model achieves at least 95% accuracy predicting the surface form of held-out test words.

In the case of Turkish, abstract URs enabled the learning of accurate morphophonological generalizations because of the pervasive amount of surface alternation. On the other hand, the lack of surface alternation in Dutch did not drive the learner to construct abstract underlying forms, yet accurate, productive morphophonological generalizations were still able to be extracted. The situation is summarized in Tab. 4.7. When generalization from concrete URs is not possible (Turkish), our model constructs abstract URs, and high generalization accuracy is achieved from these. On the other hand, when generalization from concrete URs is possible (Dutch), our model need not construct abstract URs, but high generalization accuracy is still achieved, directly from the concrete representations.

Table 4.7: The logic of model's functioning, and how high generalization accuracy is achievable both when abstraction is needed and when it is not.

Language	Concrete Generalization?	Abstract URs	High Accuracy
Turkish	✗	✓	✓
Dutch	✓	✗	✓

4.6.1 Limitations and Future Directions

The results in this chapter are promising, but more work is needed on the problem of underlying forms, and future work will need to evaluate the model on other languages.

We identified two stages at which underlying abstraction can become necessary. First, the number of forms that a morpheme takes across word forms can become large enough that underlying abstraction is necessary to predict its surface form across these words. Second, alternations across a set of morphemes can accumulate so as to prevent the formulation of productive morphological inflection rules. The first was prevalent enough in the heavily-agglutinating language Turkish to motivate underlying abstraction, but neither force necessitated abstraction in the case of Dutch nouns. An important direction for future work is to figure out how these two possible reasons for abstraction relate, including whether they can be combined.

Furthermore, we have not yet considered the question of *degrees* of abstractness. For instance, how can we provide a learning-based decision for whether to treat a Turkish alternating affix vowel as underspecified /A/ derived into [ɑ] and [e] via a harmony rule, versus /a/ derived into [e] or /e/ into [a]?

Lastly, we have assumed to this point that morphological segmentations are available to the learner. As discussed in § 4.2.1, there is experimental evidence that children are able to perform morphological segmentation. However, the task of constructing URs is intertwined with the segmentation problem, as one cannot hope to determine the UR of a morpheme without at least knowing which segments belong to its own surface realizations and not those of another morpheme. Thus, in future work we will pursue a learning-based account of segmentation and attempt to bring the problems together, jointly segmenting surface forms, learning underlying forms, and morphophonological grammars.

CHAPTER 5

An Algorithmic Account of Phonological Rules

The material in this chapter is derived from the paper "A Learning-Based Account of Local Phonological Processes" (Belth, 2023a), which has been accepted for publication at the journal ***Phonology****.*

When abstract underlying forms are constructed, phonological theory has viewed rules or constraints as mapping between these underlying forms and their concrete realizations. This chapter provides an algorithmic account of how learners may construct phonological rules. Our proposed learning algorithm models the learner's attention as initially fixed locally and expanding farther only when local dependencies do not suffice. The proposed model fits within the view of learning that we argued for in chapters § 1-2, where learning of a phonological process is triggered when, and only when, underlying abstraction introduces discrepancies between underlying and surface representations. Our proposed model successfully learns local phonological generalizations, including those lacking substantive phonetic motivation, and combines these into an ordered list of rules. The model's learned rules achieve high accuracy on held-out test words, demonstrating its efficiency learning with small data. Moreover the model's rule-construction strategy—starting as locally as possible and expanding outward—leads it to accurately exhibit the same preference for local patterns that humans do on existing experiments and on our own experiment in chapter § 6.

5.1 Introduction

Phonological processes tend overwhelmingly to involve dependencies between adjacent segments (Gafos, 2014; Chandlee et al., 2014). For example, the English plural allomorph depends on the stem-final segment, to which it is adjacent, as in (89).

(89) /dɑg-z/ → [dɑgz]
/kæt-z/ → [kæts]
/hɔrs-z/ → [hɔrsəz]

Moreover, underlying forms are often considered to be minimally different from surface forms, only exhibiting abstractness when surface alternation necessitates it (Kiparsky, 1968; Peperkamp et al., 2006; Ringe and Eska, 2013; Richter, 2021). This is supported by experimental findings, where children avoid introducing discrepancies between surface and underlying forms when there is little motivation for doing so (Jusczyk et al., 2002; Coetzee, 2009; Kerkhoff, 2007; Van de Vijver and Baer-Henney, 2014).[1]

When—and only when—concrete representations are abandoned in favor of (minimally) abstract underlying representations, a child must learn a phonological process to derive the surface form from the abstract underlying form. Experimental studies are revealing about the mechanism underlying sequence learning: humans show a strong proclivity for tracking adjacent dependencies, only beginning to track non-adjacent dependencies when the data overwhelmingly demands it (Saffran et al., 1996, 1997; Aslin et al., 1998; Santelmann and Jusczyk, 1998; Gómez, 2002; Newport and Aslin, 2004; Gómez and Maye, 2005).[2] As Gómez and Maye (2005, p. 199) put it, 'It is as if learners are attracted by adjacent probabilities long past the point that such structure is useful.' Indeed, artificial language experiments have repeatedly demonstrated that learners more easily learn local phonological processes than non-local ones (Baer-Henney and van de Vijver, 2012) and, when multiple possible phonological generalizations are consistent with exposure data, learners systematically construct the most local generalization (Finley, 2011; White et al., 2018; McMullin and Hansson, 2019).

In this chapter, we hypothesize a mechanistic account of how learners construct phonological generalizations, modeling the learner's attention as initially fixed locally and expanding farther only when local dependencies do not suffice. Our proposed model incorporates the idea that the learning of a phonological process is triggered when, and only when, underlying abstraction introduces discrepancies between underlying and surface representations (Kiparsky, 1968). We view the model's locally-centered attention and default identity assumption as being computationally parsimonious, and thus call it the *Parsimonious Local Phonology* learner (PLP). When presented with small amounts of child-directed speech, PLP successfully learns local phonological generalizations. PLP's search strategy—starting as locally as possible—leads it to accurately exhibit the same preference for local patterns that humans do. Next we review experimental results on locality (§ 5.1.1), the view of learning that PLP adopts § 5.1.2, and how these reflect principles of efficient computation § 5.1.3. See sections § 2.3 and § 2.4 for further discussions of these points.

[1] See § 2.3 for more discussion of these results.

[2] See § 2.4 for more discussion of these results.

5.1.1 Locality

Early studies of statistical sequence learning found infants to only be sensitive to dependencies between adjacent elements in a sequence. Saffran et al. (1996, 1997) and Aslin et al. (1998) found infants as young as 8-months old to be sensitive to dependencies between *adjacent* elements, but Santelmann and Jusczyk (1998) found that even at 15-months-old, children did not track dependencies between *non-adjacent* elements. Studies with older participants revealed that the ability to track non-adjacent dependencies does eventually emerge: adults show a sensitivity to dependencies between non-adjacent phonological segments (Newport and Aslin, 2004), and 18-month-old children can track dependencies between non-adjacent morphemes (Santelmann and Jusczyk, 1998). However, even as sensitivity to non-adjacent dependencies develops, learners still more readily track local dependencies. Gómez (2002) found that 18-month-olds could track non-adjacent dependencies, but that they only did so when adjacent dependencies were unavailable. Gómez and Maye (2005) replicated these results with 17-month-olds, and attempted to map the developmental trajectory of this ability to track non-adjacent dependencies, finding that it grew gradually with age. At 12 months, infants did not track non-adjacent dependencies, but they began to by 15 months, and showed further advancement at 17 months. These experiments involved a range of elements: words, syllables, morphemes, phonological segments. Moreover, similar results have been observed in different domains, such as vision (Fiser and Aslin, 2002). Together, these results suggest that learners might only discover local patterns at early stages in development, and that even after sensitivity to less-local patterns emerges, a preference for local patterns persists.

Further experiments targeted phonological learning in particular. Subjects in Finley (2011)'s artificial language experiments learned bounded (local) harmony patterns and did not extend them to non-local contexts when there is no evidence for it. However, when exposed to unbounded (non-local) harmony patterns, subjects readily extended them to local contexts. This asymmetry suggests that learners will not posit less-local generalizations until the evidence requires it. In a different study, McMullin and Hansson (2019) found that these results replicate with patterns involving liquids and with dissimilation. Baer-Henney and van de Vijver (2012) used an artificial language experiment to test the role of locality (as well as substance and amount of exposure) in learning contextually-determined allomorphs. They found that when the allomorph was determined by a segment two positions away, learners more easily acquired and extended the pattern than when the allomorph was determined by a segment three positions away. In short, these studies demonstrate that learners posit the most local generalization consistent with the data.

5.1.2 The Nature of the Learning Task

We adopt the view of others (e.g., Hale and Reiss 2008; Ringe and Eska 2013; Richter 2021) that children initially store words concretely, and only posit abstract underlying forms when motivated to do so by surface alternation. We reviewed evidence for this in § 2.3. The positing of an abstract underlying form introduces discrepancies between underlying and surface forms. For example, as Richter (2018, 2021) characterized in detail, alternations such as 'eat' [it] ~ 'eating' [iɾɪŋ] lead to the flap [ɾ] and stop [t] being collapsed into allophones of underlying /T/. Similarly, a morphemic surface alternation such as 'cats' [kæts] ~ 'dogs' [dɑgz] may motivate an abstract underlying plural suffix /-Z/ (or default /-z/) (Berko, 1958). This view is in the spirit of Kiparsky (1968)'s *alternation condition*, and has been termed *invariant transparency* (Ringe and Eska, 2013).

A consequence is that when, and only when, concrete segments are collapsed into abstract underlying representations, the need for a phonological grammar arises, to derive the surface form for abstract underlying forms. We will use the example of stops following nasals to exemplify two significant corollaries. Voiceless stops following nasals are often considered to be a marked sequence, because post-nasal articulation promotes voicing and post-nasal voicing is typologically pervasive (Locke, 1983; Rosenthall, 1989; Pater, 1999; Hayes and Stivers, 2000; Beguš, 2016, 2019). Nevertheless, many languages—e.g. English—tolerate post-nasal voiceless stops,[3] and a few even exhibit productive, phonological post-nasal *devoicing*. For example, Coetzee and Pretorius (2010) performed a detailed experimental study of Tswana speakers, finding that some extended post-nasal devoicing, as in (90; data from Coetzee and Pretorius 2010, p. 406), productively to nonce words.

(90)
a.	/m-batla/	→ [mpatla]	'want me'
	/m-botsa/	→ [mpotsa]	'ask me'
	/m-bulela/	→ [mpulela]	'open (for) me'
b.	/re-batla/	→ [rebatla]	'want us'
	/re-botsa/	→ [rebotsa]	'ask us'
	/re-bulela/	→ [rebulela]	'open (for) us'

Beguš (2019, p. 699) found post-nasal devoicing to be reported as a sound change in thirteen languages and dialects, and argued that despite appearing to operate against phonetic motivation, it likely emerged in each case as the result of a sequence of individually phonetically-motivated sound changes.

Including a constraint to mark post-nasal voiceless stops in languages that tolerate them makes the learning task unnecessarily difficult, because the constraint must then be downranked despite there being no surface alternation present. Instead, under invariant transparency, children learning

[3]We note that passive, phonetic post-nasal voicing still occurs in some such languages (Hayes and Stivers, 2000); we are referring here to phonological voicing.

languages that tolerate post-nasal voiceless stops will simply not learn a phonological process regarding post-nasal stops, because there is nothing to learn. Moreover, when surface alternations that lack or operate in opposition to phonetic motivation (e.g., post-nasal devoicing) occur synchronically due to diachronic processes or other causes, no serious problem arises: the child simply learns a phonological process to account for the observed alternation, as has been observed in experiments (Seidl and Buckley, 2005; Beguš, 2018).

The view that children initially hypothesize identity between surface and underlying forms enjoys experimental support. Jusczyk et al. (2002) found that 10-month-old infants better recognize faithful word constructions than unfaithful ones. Van de Vijver and Baer-Henney (2014) found that both 5-7yr-olds and adults were reluctant to extend German alternations to nonce words, preferring instead to treat the nonce SRs as identical to their URs. Kerkhoff (2007) reports a consistent preference for non-alternation in Dutch children ages 3-7yrs. In an artificial language experiment, Coetzee (2009) found that learners more often extend non-alternation than alternation to test words, suggesting that this is learners' default.

Of course, children's initial productions are not faithful to adult productions (Smith et al., 1973; Fikkert, 1994; Grijzenhout and Joppen, 1998; Grijzenhout and Joppen-Hellwig, 2002; Freitas, 2003), but this is likely due to underdeveloped control of the child's articulatory system, rather than an early state of the adult grammar (see Hale and Reiss 2008, sec 3.1 for a detailed argument). For instance, children systematically fail to produce complex CC syllable onsets in early speech even in languages that allow complex onsets, like Dutch, German, Portuguese, and English (Fikkert, 1994; Grijzenhout and Joppen, 1998; Grijzenhout and Joppen-Hellwig, 2002; Freitas, 2003; Gnanadesikan, 2004). Clusters tend to be reduced by deleting a consonant, and development proceeds from a cluster reduction stage to a full CC production stage, suggesting the discrepancy may be due to limited articulatory control.

PLP is a model of how phonological processes are learned once underlying abstraction leads to discrepancies in (UR, SR) pairs, which constitute PLP's input. As some of the reviewers of our Belth (2023a) paper pointed out, the task of learning phonological processes to account for discrepancies between underlying and surface forms is intertwined with the task of figuring out when such abstract underlying representations are formed, and what they are like. This is evident when comparing the English PL voicing alternation (e.g. 'cats' [kæts] ~ 'dogs' [dɑgz]) to the Dutch PL voicing alternation (e.g. 'bed' [bɛt] ~ 'beds' [bɛdən]) (Kerkhoff, 2007, p. 1). English speakers show clear productive, rule-like behavior (Berko, 1958), while Dutch speakers' generalization is less-clearly rule-like (Ernestus and Baayen, 2003; Kerkhoff, 2007). The Dutch alternation is obfuscated by its interaction with other voicing alternations, such as assimilation (Buckler and Fikkert, 2016, sec. 2). Consequently, it may be that the English alternation is systematic enough to drive the learner to systematic underlying abstraction, while the Dutch alternation is not.

Thus, a complete theory of phonological learning must include, in addition to the mechanism by which processes are learned, a precise mechanism characterizing how and when abstract underlying forms are posited. For example, Richter (2018, 2021) has hypothesized a mechanism by which learners abandon the null hypothesis of concrete underlying forms in favor of abstraction, and applied it to the case of the English [t] / [ɾ] allophones. The results closely matched lexical studies of child utterances, including a U-shaped development curve. Thus, PLP is just one part of the story. However, we believe that this part of the story—learning phonological processes from (UR, SR) pairs—is nevertheless important, and in line with the vast majority of prior work on learning phonological grammars, which have likewise tended to presuppose abstract underlying forms for use in, for example, constraint ranking (Tesar and Smolensky, 1998; Legendre et al., 1990; Boersma, 1997; Smolensky and Legendre, 2006; Boersma and Hayes, 2001; Boersma and Pater, 2008).[4] Chapter § 4 takes steps towards providing an algorithmic, learning-based account of underlying forms. In § 7.3.1 we suggest how these components may be put together.

5.1.3 Locality and Identity as Principles of Computational Efficiency

Locality and identity have natural interpretations as principles of computational efficiency, or "third factors" (Chomsky, 2005; Yang et al., 2017). The more local the context around an underlying segment, the fewer segments the cognitive system must be sensitive to (Rogers et al., 2013, p. 99) in determining its output. Moreover, it is computationally simpler to copy input segments *unaltered* to the output than to change them in the process.

We present our proposed model in § 5.2, discuss prior models in § 5.3, evaluate the model in § 5.4, and conclude with a discussion in § 5.5.

5.2 Model: PLP

Our proposed model is called the *Parsimonius Local Phonology* learner (PLP). PLP learns from an input of (UR, SR) pairs, which may grow over time as the learner's vocabulary expands. It constructs the generalizations necessary to account for which segments surface unfaithfully in those pairs and in what phonological contexts that happens. These generalizations are placed in a grammar, for use in producing output SRs for input URs.

(91) **Input:** (UR, SR) pairs

1. Initialize an empty grammar G and empty vocabulary V

[4] One reviewer of our Belth (2023a) paper pointed out that the concept of underlying forms faces skepticism, and that many phonologists have rejected the concept all together. We acknowledge that the view of learning described here is not uncontroversial. Hyman (2018) provides a discussion of the merits of underlying representations.

2. **While** there are more pairs (u, s) to learn from **do**
3. – Update V with (u, s)
4. – Use G to predict surface representation $\hat{s}$ for underlying u (§ 5.2.3.4)
5. – **For** each discrepancy between u and s not accounted for in $\hat{s}$ **do** (§ 5.2.1)
6. — Construct a generalization g for the discrepancy (§ 5.2.2)
7. — Encode g in G (§ 5.2.3)
8. – Update any generalizations that now overextend due to V growth (§ 5.2.4)

PLP assumes identity between URs and SRs by default via the fact that it only adds generalizations to G at steps 6-7, when discrepancies arise. A locality preference emerges via the generalization strategy it employs in steps 6 and 8: PLP starts with the narrowest context around an unfaithfully-surfacing segment and proceeds further from the segment only when an adequate generalization can not be found. Consequently, we consider steps 6 and 8, together with the addition of generalizations to the grammar only when motivated by discrepancies, to be PLP's main contributions. The code is available on Github.[5]

5.2.1 The Input

The input to PLP is set of (UR, SR) pairs, which may grow over time, simulating the learner's vocabulary growth. As discussed in § 5.1.2, discrepancies between a UR and its corresponding SR arise when a learner abandons concrete underlying representations in favor of underlying abstraction. A discrepancy can be an input segment that does not surface (deletion), an output segment that has no input correspondent (epenthesis), or an input segment with a non-identical output correspondant (segment change). In this work, we treat the (UR, SR) pairs, with discrepancies present, as PLP's input. Chapter § 4 and future work combine this with the important problem of when abstract underlying forms are posited (e.g., Richter 2018). We also assume that the correspondence between input and output segments is known. The same assumption is tacit in constraint ranking models, which use the correspondence for computing faithfulness constraint violations.

The URs and SRs are sequences of segments, which we treat as sets of distinctive features (Jakobson and Halle, 1956; Chomsky and Halle, 1968). Thus, structuring sound into a phonological segment inventory organized by distinctive features is treated as a separate learning process (e.g. Mayer 2020). We use feature assignments from Mortensen et al. (2016).

We will use the English plural allomorph as a running example. Suppose that at an early stage in acquisition, a child has memorized some of the plural forms of nouns in their vocabulary, as shown in (92).

[5] https://github.com/cbelth/PLP

(92) /dɑgz/
/kæts/
/hɔrsəz/
⋮

At this stage, an empty grammar, which regurgitates each memorized word, will suffice. Moreover, since no discrepancies yet exist, PLP will be content with this empty grammar: the for loop (91; step 5) will not be entered. As the child begins to learn morphology, they may discover the morphological generalization that plurals tend to be formed by suffixing a /-z/. All of the child's plural URs will then, in effect, be reorganized as in (93).

(93) /dɑg-z/
/kæt-z/
/hɔrs-z/
⋮

At this point, when the child goes to use their grammar (91; step 4), they will discover that it now predicts *[kætz] and *[hɔrsz], inconsistent with their expectation based on prior experience with the words. The newly-introduced discrepancies trigger the for loop (91; step 5) and require PLP to provide an explanation for them. Suppose the first word to trigger this is /kæt-z/, erroneously predicted as *[kætz] instead of the expected [kæts]. PLP then constructs a generalization to capture the phonological context in which /z/ surfaces as [s] (91; step 6).

5.2.2 Constructing Generalizations

The core component of PLP is its component for constructing generalizations (91; step 6).

5.2.2.1 The Structure of Generalizations

The generalizations that PLP constructs are pairs $g = (\bar{s}, c) \in \mathcal{S} \times \mathcal{A}$, where $\bar{s} \in \mathcal{S}$ (94) is a sequence and $c \in \mathcal{A}$ (95) is an action carried out at a particular position in the sequence. Each element in a sequence is a set of segments from the learner's segment inventory, Σ (94).[6]

(94) $\mathcal{S} \triangleq \bigcup_{k=1}^{\infty} \{s_1 s_2 ... s_k : s_i \subset \Sigma\}$

[6]We allow these elements to also contain syllable/word-boundary information, which we implement following Chomsky and Halle (1968); Hayes and Wilson (2008) by introducing a [±segment] feature and corresponding −segment element in Σ to mark boundaries.

A set of segments may be extensional, e.g., $s_i = \{\text{s}, \text{ʃ}, \text{z}, \text{ʒ}\}$, or a natural class—e.g., s_i = [+sib]. An action can be any in (95): deletion of the ith segment, insertion of new segment(s) to the right of the ith segment,[7] or setting the ith segment's feature f to '+' or '−'.[8]

(95) $\mathcal{A} \triangleq \{\textsc{Del}(i), \textsc{Ins}(s_{\text{new}}, i), \textsc{Set}(f, \pm, i)\}$

For example, the generalization (96a) states that a consonant is deleted when it follows and precedes other consonants, (96b) states that a 'ə' is inserted to the right of any sibilant that precedes another sibilant, and (96c) states that the voicing feature of voiced obstruents in syllable final position is set to '−' (we use '$]_\sigma$' $\in \Sigma$ to mark syllable boundary).

(96) a. ([+cons][+cons][+cons], Del(2))

b. ([+sib][+sib], Ins('ə', 1))

c. ([+voi,−son]$]_\sigma$, Set(voi, '−', 1))

Any grammatical formalism capable of encoding these generalizations could be used, but in this chapter we chose a rule-based grammar. The specified set of possible actions is meant to cover a majority of phonological processes, but more could be added if necessary (e.g. metathesis).

The part of the sequence picked out by the action's index i determines the rule's target, and the part of the sequence to the left and right of i determine the rule's left and right contexts. Each type of action (95) can be encoded in one of the rule-schemas in (97), where $k = |\bar{s}|$.

(97)

$\textsc{Del}(i)$	$s_i \rightarrow \emptyset \;/\; s_1 \ldots s_{i-1} \_\_ s_{i+1} \ldots s_k$
$\textsc{Ins}(s_{\text{new}}, i)$	$\emptyset \rightarrow s_{\text{new}} \;/\; s_1 \ldots s_i \_\_ s_{i+1} \ldots s_k$
$\textsc{Set}(f, \text{'+'}, i)$	$s_i \rightarrow [+f] \;/\; s_1 \ldots s_{i-1} \_\_ s_{i+1} \ldots s_k$
$\textsc{Set}(f, \text{'−'}, i)$	$s_i \rightarrow [-f] \;/\; s_1 \ldots s_{i-1} \_\_ s_{i+1} \ldots s_k$

Thus, the generalizations in (96) are encoded as the rules in (98).

(98) a. [+cons] → ∅ / [+cons] __ [+cons]

b. ∅ → ə / [+sib] __ [+sib]

c. [+voi, −son] → [−voi] / __ $]_\sigma$

Each sequence in $\mathcal{S}$ is *strictly local* (McNaughton and Papert, 1971)—describing a contiguous sequence of segments (cf. § B.1)—and has the same structure as the 'sequence of feature matrices' constraints from Hayes and Wilson (2008, p. 391). Moreover, the input-output relations described

[7] Insertion in initial position is achieved with $i = 0$.

[8] More generally, we can treat the first parameter as a vector of features and the second as a vector of ± values to capture multiple feature changes, but for simplicity we only describe the case of a single feature change.

by each generalization are probably[9] *input strictly local* maps (Chandlee, 2014). These structures are not necessarily capable of capturing *all* phonological generalizations, and intentionally so. Typological considerations point to strict locality as a central property of generalizations, due to its prevalence (Chandlee, 2014) and repeated occurrence across representations (Heinz et al., 2011). This paper is intentionally targeting precisely those generalizations, and we discuss principled extensions for non-local generalizations in § 5.5.1.

5.2.2.2 Searching Generalizations

When PLP encounters a discrepancy—an input segment surfacing unfaithfully—it uses algorithm (99) to construct a generalization $g = (\bar{s}, c)$. We refer to the discrepancy as $x \rightarrow y$, where x is the input segment and $y \neq x$ is what it surfaced as.

(99) **Input:** A discrepancy, $x \rightarrow y$, and the current training vocabulary V

1. Initialize a window $\bar{w} = [\{x\}]$ of width one
2. Infer c from $x \rightarrow y$ and initialize a generalization $g = (\bar{w}, c)$
3. **While** g is insufficiently accurate over V **do**[10] (§ 5.2.2.3)
4. – Expand the width of the window by length one (§ 5.2.2.4)
5. – Set g's sequence $\bar{s}$ to the most accurate context around x that fits in $\bar{w}$ (§ 5.2.2.4)

PLP uses a window, $\bar{w}$, to control the breadth of its search. The window is a sequence of cells that can be filled in to create g's sequence (94). The window starts with only one cell, filled with $s_0 = \{x\}$ (99; step 1). PLP then infers the type of change from $x \rightarrow y$ (100).

(100)

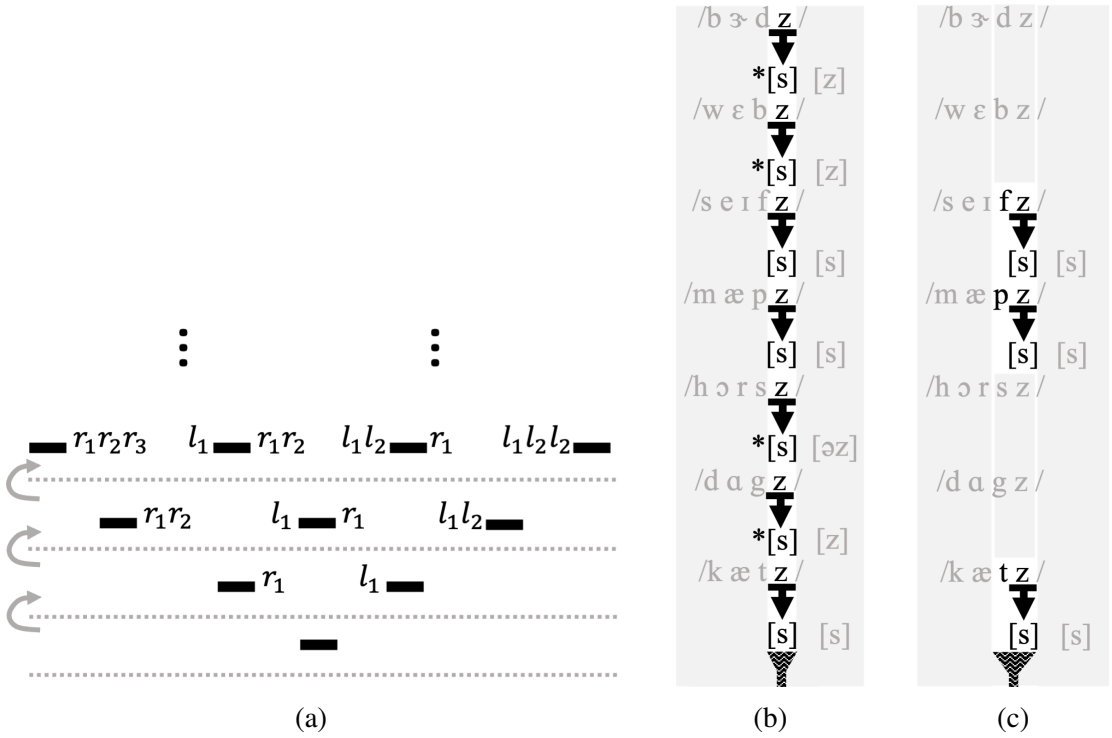
$$a = \begin{cases} \textsc{Del} & \text{if } x \rightarrow \emptyset \;\; (y = \emptyset) \\ \textsc{Ins} & \text{if } \emptyset \rightarrow y \;\; (x = \emptyset) \\ \textsc{Set} & \text{otherwise} \end{cases}$$

For Ins, the value inserted (s_{new}) is y; for Set, the featured changed (f) and its value ($\pm$) are inferred from the difference between x and y. The index, i, specifies where x falls in g's sequence, $\bar{s}$; initially since $\bar{s} = \bar{w} = [\{x\}]$, $i = 1$ (99; step 2).

As Fig. 5.1a visualizes, PLP starts with the most local generalization, which makes no reference to the segment's context: the segment always surfaces unfaithfully. In the running example, PLP first posits (101), which predicts that /z/ always surfaces as [s] (Fig. 5.1b).

[9] It is generally believed that processes describeable with the types of rules that PLP constructs are input strictly local maps (Chandlee, 2014), but—to our knowledge—there does not exist a published proof of this fact. See § B.1 for more.

[10] The loop also exits if the search runs out of context, in which case no sufficiently accurate generalization is possible.



(a) (b) (c)

Figure 5.1: **(a)**: The width of PLP's search is expanded (upward arrows) when and only when an adequate generalization cannot be found in virtue of a less-wide context. **(b)**-**(c)**: An example of PLP's search: the first generalization (101) fails because it makes too many wrong predictions, but the second (103a) allows the /z/ → [s] instances to be isolated.

(101) /z/ → [−voi] / __

This, however, is contradicted by other words in the vocabulary: /z/ surfaces faithfully as [z] in words like [dɑgz] and with an epenthetic vowel in words like [hɔrsəz], which suggests that this initial generalization is wrong (99; step 3) and that the breadth of the search must be expanded (Fig. 5.1a).

5.2.2.3 When to Expand Breadth of Search

To come to such a verdict, PLP computes the number of predictions the rule makes over the current vocabulary and how many of those were correct. The number of predictions (101) makes is the number of times /z/ appears in the learner's vocabulary, and those that surface as [s] are the correct predictions. There are a number of options for determining the adequacy of the generalization. We could require a perfect prediction record, but this may be too rigid due to the near inevitability of exceptions in naturalistic data. More generally we could place a threshold on the number or fraction of errors that the generalization can make. The choice of criterion does not substantially

change PLP: proceeding from local generalizations to less-local generalizations proceeds in the same way regardless of the quality criterion, which simply determines the rate at which the more local generalizations are abandoned. In this work, PLP uses the Tolerance Principle (Yang, 2016) as the threshold, which states that a generalization making n predictions about what an underlying segment surfaces as is productive—and hence the while loop (99; step 3) can be exited—if and only if the number of incorrect predictions it makes (called e for *exceptions*) satisfies (102).

(102)
$$e \leq \frac{n}{\ln n}$$

The threshold is cognitively motivated, predicting that children accept a linguistic generalization when it is cognitively more efficient to do so (see Yang 2016, ch. 3 for the threshold's derivation). Since the threshold is based on cognitive considerations and has had success in prior work (e.g. Schuler et al. 2016; Koulaguina and Shi 2019a; Emond and Shi 2021; Richter 2021; Belth et al. 2021), it is a reasonable choice for this chapter (see § 2.5 for a discussion of the Tolerance Principle). In the current example, (101) has $n = 7$ and $e = 4$, which fails to pass (102): $4 > 7/\ln 7$. Thus, the while loop (99; step 3) is entered.

5.2.2.4 Expanding Breadth of Search

Once the initial hypothesis that /z/ always surfaces as [s] is ruled out as too errant, PLP adds one cell to the window (99; step 4). PLP fills the window with the sequence that matches the fewest of the sequences where /z/ does not surface as [s]. In other words, it chooses the context that better separates words like /kæt/ from words like /dɑg/ and /hɔrs/. Thus, for the vocabulary in Fig. 5.1b-5.1c, PLP prefers (103a) over (103b)[11] because a left context of {t, p, f} better separates the places where /z/ does indeed surface as [s] from those where it does not, than does a right context of {#}. That is, PLP chooses the rule with the most accurate context fitting in the current window, where accuracy is measured as the fraction of the rule's predictions over the training URs that match the corresponding training SRs. In our example then, PLP's second hypothesis is that /z/ surfaces as [s] whenever it follows a /t/, /p/, or /f/.

(103) a. /z/ → [−voi] / {t, p, f} __

b. /z/ → [−voi] / __ {#}

Figure 5.1a visualizes PLP's search: it hypothesizes a context where an underlying segment surfaces as some particular segment other than itself, checking whether the hypothesis is satisfactorily accurate, and expanding the breadth of its search if not. This process halts once a sufficiently accurate hypothesis has been discovered.

[11]The symbol '#' denotes a word boundary.

5.2.3 Encoding Generalizations in a Grammar

The generalizations that PLP constructs are encoded in a grammar to be used in producing an SR for an input UR. The grammar, G, consists of a list of rules. Each time PLP constructs a generalization (91; step 6), it is placed in the appropriate rule schema (97) and added to the list of rules. If PLP replaces a generalization due to underextension or overextension (91; step 8), as described in § 5.2.4, the old, offending rule is removed and a new one added. § 5.2.3.1 discusses how rules that carry out the same action are combined, § 5.2.3.3 discusses how the list of rules is ordered, § 5.2.3.2 discusses how natural classes are induced, and § 5.2.3.4 discusses how G produces outputs from inputs.

5.2.3.1 Combining Generalizations

Generalizations that carry out the same change over different segments are combined in the grammar, so long as the resulting rule is satisfactorily accurate, via (102). For instance, the three rules in (104a) would be grouped into the single rule (104b).

(104) a. /d/ → [−voi] / __]$_\sigma$
/v/ → [−voi] / __]$_\sigma$
/g/ → [−voi] / __]$_\sigma$
b. {d, v, g} → [−voi] / __]$_\sigma$

5.2.3.2 Inducing Natural Classes

Up to this point, PLP's generalizations have been over sets of particular segments. Humans appear to generalize from individual segments to natural classes, as has been recognized by theory (Chomsky and Halle, 1965; Halle, 1978; Albright, 2009) and evidenced by experiment (Berent et al., 2007; Finley and Badecker, 2009; Berent, 2013).

PLP thus attempts to generalize to natural classes for each set of segments in a generalization's sequence $\bar{s}$, in terms of shared distinctive features (Jakobson and Halle, 1956; Chomsky and Halle, 1968). The procedure can be thought of as retaining only the features shared by segments in $\bar{s}$ needed to keep the rule satisfactorily accurate. To exemplify this part of the model, we will assume PLP has constructed the epenthesis rule (105)—e.g., /hɔrsz/ → [hɔrsəz].

(105) ∅ → ə / {s, ʃ, z} __ {z}

The procedure, outlined in (106), starts with a new length-$|\bar{s}|$ sequence $\bar{n}$, with each element an empty natural class (106; step 1).

(106) **Input:** A generalization $g = (\bar{s}, c)$

1. Initialize a new generalization $g_{\text{nc}} = (\bar{n}, c)$ with empty natural classes, $\bar{n}$
2. Initialize feature options for natural classes
3. **While** g_{nc} is insufficiently accurate over V **do**
4. – Add to $\bar{n}$ the feature that best narrows $\bar{n}$'s extension down to $\bar{s}$'s
5. Replace g with g_{nc}

For generalization (105), the sequence $\bar{s}$ is (107a) and the (empty) initial natural class sequence is (107b). Each element of $\bar{n}$ can take any feature shared by the corresponding segments in $\bar{s}$, so the set of feature options is (107c), which includes elements like (+sib, 1) because {s, ʃ, z} share '+sib' as a feature and (+voi, 2) because {z} has '+voi' as a feature, but it does not include (+voi, 1) because {s, ʃ, z} do not agree on this feature.

(107) a. $\bar{s}$ = {s, ʃ, z}{z}

b. $\bar{n}$ = [][]

c. {(+cons, 1), (+sib, 1), (−son, 1), ..., } ∪ {(+sib, 2), (+voi, 2) ...}

Inside the while loop (106; step 3), features are added one at a time to $\bar{n}$, choosing at each step the feature from (107c) that best narrows the extension of $\bar{n}$ (initially all length-$|\bar{s}|$ sequences) to those in the extension of $\bar{s}$ (which is {sz, ʃz, zz}). Thus, adding the feature '+sib' to the first natural class (108a) will narrow $\bar{n}$'s extension towards $\bar{s}$'s better than '+cons.' The new generalization, g_{nc} is evaluated as before with Tolerance Principle, via (102). In the current example, $\bar{n}$ (108a) will still have sequences like {st, zi, ʃu, ...} in its extension, so '+sib' will then be added to the second natural class (108b).

(108) a. $\bar{n}$ = [+sib][]

b. $\bar{n}$ = [+sib][+sib]

This new sequence, $\bar{n}$, still has an extension greater than the original $\bar{s}$. However, because adjacent sibilants are indeed disallowed in English, this inductive leap is possible, and thus (105) will be replaced with (109) in the grammar.

(109) $\emptyset \rightarrow$ ə / [+sib] __ [+sib]

This differs from the natural class induction in Albright and Hayes (2002, 2003), which generalizes as conservatively as possible by retaining all shared features (see § B.2.3).

It may be possible for natural class induction to influence rule-ordering, so PLP induces them prior to rule ordering. Specifically, natural classes are induced with rules temporarily ordered by scope (narrowest first), before the final ordering is computed as in § 5.2.3.3.

5.2.3.3 Rule Ordering

In some cases, phonological processes may interact, in which case the interacting rules may need to be ordered. The topic of rule interaction and ordering has received immense attention in the literature—especially in discussions of opacity—and is well-beyond the scope of the current chapter to fully take up here. However, we will summarize PLP's approach to rule ordering, and characterize the path to a more systematic study of PLP's handling of complex rule interactions.

The standard rule interactions discussed in the literature are FEEDING, BLEEDING, COUNTERFEEDING, and COUNTERBLEEDING, described in (110) following McCarthy (2007); Baković (2011).

(110) Given two rules r_i and r_j, where r_i precedes r_j,

a. r_i FEEDS r_j iff r_i creates additional inputs to r_j

b. r_i BLEEDS r_j iff r_i destroys potential inputs to r_j

c. r_j COUNTERFEEDS r_i iff r_j creates additional inputs to r_i

d. r_j COUNTERBLEEDS r_i iff r_j destroys additional inputs to r_i

COUNTERFEEDING and COUNTERBLEEDING are *counterfactual inverses* of FEEDING and BLEEDING: if r_j COUNTERFEEDS (resp. COUNTERBLEEDS) r_i, it would FEED (resp. BLEED) r_i if it preceded r_i. McCarthy (2007, sec. 5.3)'s example of FEEDING, reproduced in (111), comes from Classical Arabic, where vowel epenthesis before word-initial consonant clusters (r_i) feeds [ʔ] epenthesis before syllable-initial vowels (r_j).

(111) /dˤrib/ (underlying) →
idˤrib (vowel epenthesis) →
ʔidˤrib ([ʔ] epenthesis) →
[ʔidˤrib] (surface) 'beat! MASC.SG.'

McCarthy (2007, sec 5.4) also provides an example of COUNTERFEEDING. In Bedouin Arabic, short high vowels are deleted in non-final open syllables, and /a/ is raised in the same environment. However, as (112) shows, because deletion precedes raising, the raising of the short vowel /a/ to [i] does not feed deletion.

(112) /dafaʔ/ (underlying) →
dafaʔ (no deletion) →
difaʔ (raising) →
[difaʔ] (surface) 'he pushed'

Examples of BLEEDING and COUNTERBLEEDING come from dialects of English where /t/ and /d/ are flapped—[ɾ]—between stressed and unstressed vowels, while /aɪ/ and /aʊ/ raise to [ʌɪ] and

[ʌʊ] before voiceless segments. The canonical case is COUNTERBLEEDING order, where raising occurs before underlying /t/ even when it surfaces as voiced [ɾ] on the surface (113).

(113) /raɪtɚ/ (underlying) →
rʌɪtɚ (raising) →
rʌɪɾɚ (flapping) →
[rʌɪɾɚ] (surface)

In lesser-discussed dialects of English in Ontario, Canada (Joos, 1942) and in Fort Wayne, IN (Berkson et al., 2017), the flapping of voiceless /t/ as voiced [ɾ] bleeds raising (114).

(114) /raɪtɚ/ (underlying) →
raɪɾɚ (flapping) →
raɪɾɚ (/aɪ/ raising does not apply due to voiced 'ɾ') →
[raɪɾɚ] (surface)

Given two interacting rules r_i and r_j, it is straight-forward to order them by following standard arguments. Specifically, ordering r_i before r_j (FEEDING/BLEEDING order), will produce errors on data from a language where r_j in fact precedes r_i (COUNTERFEEDING/COUNTERBLEEDING) and vice versa. For example, if we call English dialects where flapping counterbleeds raising (113) 'Dialect A' and the dialects with bleeding (114) 'Dialect B', ordering flapping before raising in 'Dialect A' will erroneously cause /raɪtɚ/ to surface as [raɪɾɚ] instead of [rʌɪɾɚ]. Consequently, the correct COUNTERFEEDING order will yield higher accuracy than FEEDING order for a leaner exposed to 'Dialect A.' A symmetrical argument holds for ordering in 'Dialect B.'

Thus, for each pair of learned rules, PLP chooses the pairwise ordering with higher accuracy. To yield a full ordering of the rules, PLP constructs a directed graph where each rule in $\mathcal{R}$ forms a node. PLP considers each pair of rules $(r_i, r_j) \in \mathcal{R} \times \mathcal{R}$ and places a directed edge from r_i to r_j iff the accuracy of $r_j \circ r_i$ (i.e., applying r_i first and r_j to its output) is greater than the reverse, $r_i \circ r_j$. The directed graph is then topologically sorted[12] to yield a full ordering. In such an ordering, the ordering between any pair of rules that do not interact is arbitrary, while that between any pair that do interact is the order that achieves higher accuracy.

The bigger challenge is the possibility that the interactions between r_i and r_j obfuscate the independent existence of the rules, thereby making it difficult for them to be discovered in the first place. COUNTERFEEDING and COUNTERBLEEDING present no issues, because applying each rule independently, directly over the UR, produces the same SR as applying them sequentially in COUNTERFEEDING/COUNTERBLEEDING order. For example, in McCarthy (2007)'s Bedouin Arabic example (112) /a/ → [i] is accounted for by the raising rule, and there is no deletion in /dafaʔ/ →

[12] A *topological sort* of a directed graph is a linear ordering of its nodes such that every ordering requirement encoded in its edges is preserved (Cormen et al., 2009, p. 612).

[difaʔ] to hinder the discovery of the deletion rule. Similarly, the /a/ → [ʌ] discrepancy in (113) can be accounted for by raising, without reference to flapping, and the /t/ → [ɾ] discrepancy can be accounted for by flapping without reference to raising. We give an empirical demonstration of PLP learning rules in COUNTERBLEEDING order in § 5.4.3.4.

Since BLEEDING destroys contexts where a rule would have applied, it can cause overextensions. For example, when PLP is attempting to construct a raising rule for (114), rule (115) (treating the diphthong as a single segment) would overextend to /raɪtɚ/.

(115) aɪ → ʌɪ / __ [−voi]

However, since PLP allows some exceptions via the Tolerance Principle, this will only matter if the bled cases are pervasive enough to push the rule over the Tolerance Principle threshold (Eq. 102). Whether this happens must be determined on a case-by-case basis by the learner's lexicon. If the threshold of exceptions is crossed, PLP will simply expand the width of its search. When flapping bleeds raising (114), raising occurs distributionally before underlying voiceless segments *that are not between a stressed and an unstressed vowel*. The latter condition describes the contexts where raising is not bled, and still falls within a fixed-size window of the raising target, as shown with underlines in /raɪtɚ/. The general point here is that if two rules interact extensively, there is likely to still be a fixed-length context—possibly a slightly larger context—that accounts for the processes. In fact, Chandlee et al. (2018) showed that a wide-range of phonological generalizations that have been characterized as opaque in the literature can be characterized as Input Strictly Local maps. In the appendix, we show that the rules PLP learns correspond to Input Strictly Local maps. Thus, we are optimistic that PLP can succeed even with instances of opaque rule interactions. § 5.4.4 provides an empirical demonstration of PLP learning rules in BLEEDING order.

FEEDING may require small adaptions to PLP. In (111), no issue arises for the vowel-epenthesis rule, which does the feeding. The search for a rule to account for epenthetic [ʔ] will proceed analogously to the BLEEDING case. There are two underlying environments where epenthetic [ʔ] surfaces: before underlyingly initial vowels (# __ V) and before underlying initial consonant clusters (i.e. where raising feeds it, # __ CC). These are disjoint contexts, so it may be appropriate to adapt PLP to allow it to return two disjoint rules from its search (99) to account for a discrepancy. In that case, the rules in (116) account for [ʔ]-epenthesis directly from URs.

(116) ∅ → ʔ / # __ CC ('fed' [ʔ]-epenthesis cases)
∅ → ʔ / # __ V

Alternatively, PLP could be adapted such that the search for new generalizations (91; step 6) operates over intermediate representations—specifically those derived by existing rules—instead of underlying representations. In that case, the [ʔ]-epenthesis rule could be directly learned over the intermediate forms derived by the vowel-epenthesis rule.

In summary, this chapter is not an attempt to provide a complete account of rule ordering, which is beyond its scope. The results in § 5.4.3.4 and § 5.4.4 provide empirical demonstration of PLP learning some interacting rules, and the above discussion provides an outline of how PLP approaches rule interaction and what extensions may be necessary. We discuss future directions for studying rule interaction in § 7.3.5.

5.2.3.4 Production

The rules are applied one after the other in the order produced by the procedure in § 5.2.3.3. Each individual rule is interpreted under *simultaneous application* (Chomsky and Halle, 1968), which means that when matching the rule's target and context, only the input is accessible, not the result of previous applications of the rule. Thus, following the example from Chandlee (2014, p. 37), the rule (117) applied simultaneously to the input string *aaaa* yields the output *abba* rather than *abaa*, because the second application's context is not obscured by the the first application.

(117) $a \rightarrow b \,/\, a \_\_ \; a$

Simultaneous application is the interpretation of rules that corresponds to input-strictly local maps, as we discuss in § B.1.2. Other types of rule application, such as iterative or directional (e.g., Howard 1972; Kenstowicz and Kisseberth 1979), could be used in future work.

Thus, for an input u and ordered list of rules $\mathcal{R} = r_1, r_2, ..., r_{|\mathcal{R}|}$, the grammar's output $\hat{s}$ is given by the composition of rules in (118).

(118) $\hat{s} = G(u) = r_{|\mathcal{R}|} \circ r_{|\mathcal{R}|-1} \circ ... \circ r_1(u) = r_{|\mathcal{R}|}(r_{|\mathcal{R}|-1}(...r_1(u)))$

5.2.4 Updating Incrementally

As PLP proceeds, vocabulary growth may cause the grammar to become stale and underextend or overextend, at which point PLP updates any problematic generalizations (91; step 8).

Denoting the discrepancies between the input u and the predicted output $\hat{s}$ as $d(u, \hat{s})$, and those between u and s as $d(u, s)$, underextensions are defined in (119a) as discrepancies between the input and expected output that are not accounted for in PLP's prediction $\hat{s}$, and overextensions are defined in (119b) as discrepancies in the predicted output that should not be there. Here '\' denotes set difference, and '$\triangleq$' means 'equal by definition.'

(119) a. $U \triangleq d(u, s) \setminus d(u, \hat{s})$

b. $O \triangleq d(u, \hat{s}) \setminus d(u, s)$

Underextensions are handled by the for loop (91; step 5). Inside the loop, a new generalization is created (91; step 6). This is encoded in the grammar (91; step 7) by adding it to this list of rules.

If a prior generalization for the discrepancy exists, it is deleted from the list. An example of this is (120), where the word /mæp-z/ (120b) freshly enters the vocabulary.

(120) a. /dɑg-z/ → [dɑgz]
/kæt-z/ → [kæts]
/hɔrs-z/ → [hɔrsəz]
b. /mæp-z/ → [mæps]

Prior to its arrival, the rule (121a) was sufficient to explain when /z/ surfaces as [s]. This, however, fails to account for the new word, which ends in /p/ not /t/. PLP handles this by discarding the old rule and replacing it with a fresh one, such as (121b), derived by the exact same process described above in § 5.2.2.

(121) a. /z/ → [−voi] / {t} __
b. /z/ → [−voi] / {t, p} __

Overextension—a discrepancy between the input u and PLP's prediction $\hat{s}$ that did not exist between u and the expected output s—is handled by (91; step 8). An example is (122), where (122b) enters the learner's vocabulary after (122a).

(122) a. /kæt-z/ → [kæts]
b. /dɑg-z/ → [dɑgz]

In such a case, the rule (123) will have been sufficient to explain (122a), but will result in an erroneous *[dɑgs] for (122b).

(123) /z/ → [−voi] / __

PLP resolves this by discarding the previous rule and replacing it with a new one via the process in § 5.2.2.

For both underextension and overextension, when the list of rules is updated, the steps in § 5.2.3—combining generalizations, inducing natural classes, and ordering rules—are repeated. Since PLP can replace generalizations as needed as the vocabulary grows, it can learn incrementally, in batches, or once and for all over a fixed vocabulary.

5.3 Prior Models

5.3.1 Constraint-Based Models

Constraint-ranking models rank a provided set of constraints. Tesar and Smolensky (1998)'s Constraint Demotion algorithm was an early constraint-ranking model for OT. Others are built on

stochastic variants of OT or Harmonic Grammar (HG) (Legendre et al., 1990; Smolensky and Legendre, 2006), including the Gradual Learning Algorithm (Boersma, 1997; Boersma and Hayes, 2001) for Stochastic OT and a later model (Boersma and Pater, 2008) that provided a different update rule for HG (see Jarosz 2019 for an overview).

Constraint ranking models can capture the assumption of classical Optimality Theory that learning amounts to ranking a universal constraint set, or they can rank a learned constraint set. Hayes and Wilson (2008)'s Maximum Entropy model learns and ranks constraints, but it learns phonotactic constraints over surface forms, not alternations as PLP does.

Locality and identity biases are better reflected in the content of the constraint set than in the constraint ranking algorithm. Locality is determined in virtue of what segments are accessed in determining constraint violations.

Constraint ranking models usually initialize markedness constraints outranking faithfulness constraints (Smolensky, 1996; Tesar and Smolensky, 1998; Jusczyk et al., 2002; Gnanadesikan, 2004). Consequently, any UR will initially undergo any changes necessary to avoid marked structures, even when lacking surface alternation to motivate discrepancies. Ranking faithfulness constraints above markedness constraints has been advocated for by Hale and Reiss (2008), but this approach has not been widely adopted. This in part due to arguments that such an initial ranking would render some grammars unlearnable (Smolensky, 1996), and in part due to the view that features of early child productions, in particular 'emergence of the unmarked,' reflect an early stage of the child's grammar, rather than underdeveloped articulatory control.

5.3.2 Rule-Based, Neural Network, and Linear Discriminative Models

Johnson (1984) proposed an algorithm for learning ordered rules from words arranged in paradigms as a proof of concept about the learnability of ordered-rule systems. The algorithm did not incorporate a locality bias and was not extensively studied empirically or theoretically.

Albright and Hayes (2002, 2003) developed a model for learning English past tense morphology via probabilistic rules. The model can be applied to learn rules for any set of input-output word pairs, including phonological rules. It is called the *Minimum Generalization Learner* because when it seeks to combine rules constructed for multiple input-output pairs, it forms the merged rule that most tightly fits the pairs. A consequence of this generalization strategy is that the phonological context of the rule is as wide as possible around the target segment, only localizing around the target when less-local (and hence less-general) contexts cannot be sustained. This is the direct opposite of PLP and of experimental results that suggest human learners start with local patterns and only move to non-local patterns when local generalizations cannot be sustained (Finley, 2011; Baer-Henney and van de Vijver, 2012; McMullin and Hansson, 2019). We further discuss differences between

PLP and MGL in § B.2.

Rasin et al. (2018) proposed a Minimum Description Length model for learning optional rules and opacity. The authors intended the model as a proof-of-concept and only evaluated it on two small, artificial datasets.

Peperkamp et al. (2006) proposed a statistical model for learning allophonic rules by finding segments with near-complementary distributions. The method is not applicable to learning rules involving non-complementary distributions. Calamaro and Jarosz (2015) extend the model to handle some cases of non-complementary distributions, if the alternation is conditioned in terms of the following segment (i.e., $a \rightarrow b / \_\_ c$ where $|a| = |b| = |c| = 1$). These works attempt to model the very early stage of learning alternations (White et al., 2008) prior to most morphological learning, whereas PLP models learning after abstract URs begin to be learned.

Beguš (2022) trained a generative, convolutional neural network on audio recordings of English-like nonce words, which followed local phonological processes and a non-local process (vowel harmony). The model was then used to generate speech. This model-generated speech followed the local processes more frequently than the non-local process, suggesting that it more easily learned local than non-local processes. This is possibly due to the use of convolution, which is a fundamentally local operation. As a model for generating artificial speech, it is not directly comparable in our setting of learning processes that map URs to SRs.

In a different direction, Baayen et al. (2018, 2019) proposed using Linear Discriminative Learning to map vector representations of form onto vector representations of meaning and vice versa. Since this model operates over vector representations of form and meaning, it is not directly comparable.

5.3.3 Formal-Language-Theoretic Models

Formal-language and automota-theoretic approaches analyze phonological generalizations in computational terms. Many resulting learning models attempt to induce a finite state transducer (FST) representation of the map between SRs and URs. These automata-theoretic models, together with precise assumptions about the data available for learning, allow for learnability results in the Gold (1967) paradigm of *identification in the limit*. Such results state that a learning algorithm will converge onto a correct FST representation of any function from a particular family, provided that the data presented to it meets certain requirements—called a *characteristic sample*. In phonology, the target class of functions is usually one that falls in the subregular hierarchy (Rogers et al., 2013), which contains classes of functions more restrictive than the *regular* region of the Chomsky Hierarchy (Chomsky, 1956). These models are often chosen to demonstrate theoretical learnability results, and have seldom been applied to naturalistic data.

Gildea and Jurafsky (1996) developed a model, based on OSTIA (Oncina et al., 1993), which learns subsequential FSTs. The class of subsequential functions is a sub-regular class of functions that may be expressive enough to capture any type of observed phonological map (Heinz, 2018), although some tonal patterns appear to be strong counter-examples (Jardine, 2016a). The authors intended their model only as a proof-of-concept of the role of learning biases, and required unrealistic quantities of data to effectively learn. Indeed, they recognized the importance of faithfulness and locality as learning biases, which they attempted to embed into OSTIA. Their biases were, however, heuristics. In particular, a bias for locality was introduced by augmenting states with the features of their neighboring contexts. This in effect restricts the learner to local patterns, which is different from the current chapter's proposal, in which locality is a consequence of the way that the algorithm proceeds over hypotheses.

As Chandlee (2014) observes, a more principled means of incorporating a locality bias into a finite state model is to directly target the class of strictly local functions. Chandlee et al. (2014) proposes such a model, called ISLFLA, and proves that it can learn any strictly local function in the limit in the sense of Gold (1967). However, the characteristic sample for the algorithm includes the set of input-output pairs for every language-theoretically possible string up to length k (a model-required parameter). As Chandlee (2014) discusses, this is problematic since natural language may in principle never provide all logically possible strings, due to phonotactic or morophological constraints. We implemented ISLFLA and attempted to run it on naturalistic data, and it does indeed fail to identify any FST on such data.[13]

Jardine et al. (2014) proposed a model, SOSFIA, for learning subsequential FSTs when the FST structure is known in advance; only the output for each arc in the FST needs to be learned. Strictly local functions are such a case, because the necessary and sufficient automata-theoretic conditions of strict locality include a complete FST structure (Chandlee, 2014). SOSFIA also admits learnability in the limit results, but has not been applied to naturalistic data.

5.4 Evaluating the Model

This section evaluates PLP along a number of dimensions (124).

(124) **Q1**. Does PLP reflect human learners' preference for local generalizations?
Q2. How well does PLP learn local generalizations?
Q3. What are the learning effects of assuming UR-SR identity by default?

[13]OSTIA will run on data not satisfying its characteristic sample; it is just not guaranteed to induce a correct FST in such cases. In contrast, ISLFLA is unable to proceed if the characteristic sample is not met: it exits at line 9 of the pseudocode in Chandlee et al. (2014, p. 499).

5.4.1 Model Comparisons

We compare to several alternative models.

5.4.1.1 Rule-Based, Neural Network, and Finite-State Models

MGL is the Minimal Generalization Learner from Albright and Hayes (2002, 2003). We used the Java implementation provided by the authors. MGL may produce multiple candidate SRs for a UR if more than one rule applies to the UR. In such cases, we used the rule with the maximum conditional probability scaled by scope (*confidence* in the terminology of Albright and Hayes 2002, sec. 3.2) to derive the predicted SR.

ED (encoder-decoder) is a neural network model. It is a successful neural network model for many natural language processing problems involving string-to-string functions, such as machine translation between languages (Sutskever et al. 2014), and morphological reinflection (Cotterell et al. 2016). It has also been used to revisit the use of neural networks in the 'past-tense debate' of English morphology (Kirov and Cotterell 2018), though its use as a computational model of morphology acquisition has been called into question (McCurdy et al. 2020; Belth et al. 2021). We follow Kirov and Cotterell (2018) and Belth et al. (2021) in its setup, using the same RNN implementation, trained for 100 epochs, with a batch size of 20, optimizing the log-likelihood of the training data. Both the encoder and the decoder are bidirectional LSTMs with 2 layers, 100 hidden units, and a vector size of 300.

OSTIA (Oncina et al., 1993) is a finite-state model for learning subsequential finite state transducers. We used the Python implementation from Aksënova (2020).

ID is a trivial baseline that simply copies every input segment to the output. This allows for interpreting the value of assuming UR-SR identity by default.

5.4.1.2 Learning as Constraint Ranking

We also compare to the view of learning as ranking a provided constraint set. Classic OT viewed constraints as part of UG; we represent this view with **UCON**, for *universal constraint set*. An alternative view is that the constraint set is learned; we represent this view with **ORACLE**, which effectively constitutes an upper-bound on how well a model that learns the constraint set to be ranked could do. ORACLE is provided all and only the markedness constraints relevant to the grammar being learned. UCON is provided the same constraints as ORACLE, plus two extra markedness constraints that are violable in the adult languages and thus must be down-ranked.

It is important to emphasize that these models learn in a different setting than PLP and those in § 5.4.1.1. The latter receive as input only UR-SR training pairs, whereas UCON and ORACLE receive both training pairs and a constraint set. Consequently, UCON and ORACLE's accuracies at

producing SRs are not directly comparable to the other models' accuracies. Our goal in comparing PLP to UCON and ORACLE is to highlight the ways in which PLP's account of phonological learning differs.

For UCON and ORACLE, we use the Gradual Learning Algorithm (GLA) (Boersma, 1997; Boersma and Hayes, 2001) to rank the constraints because it is robust to exceptions—an important property when learning from noisy, naturalistic data. We emphasize, however, that the comparison is not to the particular constraint-ranking algorithm—others could have been chosen. Because our experiments involve many random samples and tens of thousands of tokens, the implementation of GLA in Praat (Boersma et al. 1999) was not well-suited. Thus, we used our own Python implementation of GLA, with the same default parameters as in Praat (evaluation noise: 2.0, plasticity: 1.0). We initialize markedness constraints above faithfulness constraints.

5.4.2 Comparison to Humans' Preference for Locality

In an experimental study, Baer-Henney and van de Vijver (2012) found that allomorphic generalizations in an artificial language were more easily and successfully learned when the surface allomorph was determined by a segment two positions away than determined by a segment three positions away. The study involved three artificial languages in which plural nouns were formed by affixing either the vowel [-y] or [-u], which differ in backness: −back and +back, respectively. Each language involved a different phonological condition for determining which affix surfaced. Treating /-y/ as the underlying affix, the three generalizations are those in (125).

(125) a. [−back] → [+back] / [+vowel,+**back**][+cons] __

b. [−back] → [+back] / [+vowel,+**tense**][+cons] __

c. [−back] → [+back] / [+cons,+**son**][+vowel][+cons] __

All singular forms are CVC words; plurals add a vowel. The (125a) language is an example of vowel harmony, since the affix vowel assimilates in backness to the preceding vowel. The (125b) language is equally local, but lacks clear phonetic motivation, since the stem vowel's feature determining the affix' backness is [tense]. The (125c) language is both less local and phonetically unmotivated, since the backness of the vowel is determined by the initial consonant of the stem. Because all three languages have CVC stems and CVCV plurals, each pattern is strictly local, but (125a)-(125b) involve a sequence of three contiguous segments while (125c) involves four.

Since PLP starts locally around the affix when looking for an appropriate generalization, and only proceeds outward when the more local contexts become too inaccurate, we expect PLP to learn the (125a)-(125b) generalizations substantially more easily than the (125c) generalization, just as Baer-Henney and van de Vijver (2012) found for humans (**Q3**). For comparison, we use MGL,

which generalizes in roughly the opposite way: it constructs the narrowest—and hence less local—generalization. We also compare to grammars resulting from ranking three different constraint sets. The markedness constraints for (125) are (126).

(126) a. *[+vowel,+back][+cons][−back,+vowel]

b. *[+vowel,+tense][+cons][−back,+vowel]

c. *[+cons,+son][+vowel][+cons][−back,+vowel]

The first constraint set encodes the assumption of a universal constraint set containing only grounded, universal constraints by including only (126a), because it is the only generalization viewed as phonetically motivated. Secondly, we consider a constraint set containing all three markedness constraints (126a)-(126c) regardless of which language is being learned. Thirdly, we consider a constraint set containing only the constraint relevant to the language being learned.

In addition to learning the local generalization more easily than the non local one, Baer-Henney and van de Vijver (2012) also found that the phonetically motivated generalization (125a) was learned slightly more easily than (125b). The authors argued that this is evidence for substantive bias in phonological learning. However, the question of substantive bias is largely orthogonal to the current chapter, since our focus is on locality. Moreover, the performance gap between (125a) and (125b) was much smaller than the gap between them and (125c). For these reasons, we focus on the difference in models' performance on (125a)-(125b) vs. (125c) in this experiment.

5.4.2.1 Setup

Each of Baer-Henney and van de Vijver (2012)'s experiments involved presenting subjects with randomly selected singulars and plurals from the respective artificial languages. Each word was accompanied by a picture conveying the word's meaning; one item was present in the picture for singulars and multiple items for plurals. The singulars and plurals were presented independently, so the experimental setup did not separate phonological learning from learning the artificial languages' morphology and semantics. Due to this fact, the study participants likely only successfully acquired the underlying and surface representations for a subset of the exposure words, and what fraction of the exposure set they learned is entirely unknown. Since the models assume URs and SRs as training data, we factor out the fraction of the exposure set for which they have acquired URs and SRs by treating it as a free-variable that the models get to optimize over. We use the data released by Baer-Henney and van de Vijver (2012) and follow their setup to construct train (exposure) and test sets.[14] We ran each model over 100 randomized exposure sets to simulate 100 participants.

[14]Baer-Henney and van de Vijver (2012) used both high and low frequency settings, where the high frequency setting included a higher fraction of plural forms in the exposure set. Since we already treat the amount of exposure data available for learning phonology as a free variable, we followed the high-frequency setting for our experiment.

The MGL model from Albright and Hayes (2002, 2003) combines rules that target the same segment and carry out the same change to that target. For instance, if it has acquired the two word-specific rules in (127a)-(127b), it will attempt to combine them via Minimal Generalization—i.e., as conservatively as possible. The minimal generalization for (127a)-(127b) is (127c), which retains as much as possible of the original two rules. However, in the implementation of MGL from Albright and Hayes (2002, 2003), when two rules are combined, the longest substrings shared by the rules are retained—in this case /p/—but segment combination (e.g., {ʊ,o}) only proceeds *one position further*; everything else is replaced by a free variable X (see Albright and Hayes 2002, p. 60 for a complete description of this process). Thus, their implementation returns (127d), which is less conservative than the actual minimal generalization (127c).

(127) a. y → u / bʊp __ #
b. y → u / dop __ #
c. y → u / {b,d}{ʊ,o}p __ #
d. y → u / X{ʊ,o}p __ #

This issue does not arise in the original papers by Albright and Hayes (2002, 2003), because the object of study was the English past tense, in which the surface allomorph is determined by an immediately adjacent segment. Thus, any regularities beyond the adjacent segment, which their implementation would miss despite being more minimal generalizations, would be spurious regularities anyway. However, for the purposes of this experiment, the implementation is problematic. Consequently, we used our own implementation of MGL, which correctly generates the minimally general combination of rules.[15] We use the rule with the minimally general context in which /-y/ surfaces as [-u] to produce a surface form for each test instance.

5.4.2.2 Results

Fig. 5.2 shows the results. The *x*-axis of each plot is the free-variable discussed above, measuring the fraction of the exposure set that the learner successfully constructed a UR-SR pair for. The *y*-axis reports the average model performance (over the 100 simulations), for each language. The points marked with a color-coded 'X' provide the average *human* performance from Baer-Henney and van de Vijver (2012) (over the 20 participants). Since each model gets to optimize over the free variable, we select, for each respective model, the *x*-value where it best matches the human performance, averaged over all three languages. This point can be seen by where the 'X's are placed. We drew a color-coded vertical line from each human performance marker to the corresponding model performance to demonstrate the difference between each model's performance and the humans'.

[15] All other experiments involving MGL used Albright and Hayes (2002, 2003)'s implementation.

PLP (Fig. 5.2a) is the best match to the human results, learning the more local generalizations (125a)-(125b) substantially more easily than the less local generalization (125c). This is because PLP requires sufficient evidence against local generalizations (via the Tolerance Principle) before it will abandon them for less local ones (§ 5.2.2.2). This is reminiscent of Gómez and Maye (2005, p. 199)'s characterization of human learners as attending to local contexts even 'past the point that such structure is useful' before eventually moving on to less local information. In contrast, MGL (Fig. 5.2b) learns all three generalizations equally well because it constructs the most conservative—and hence widest context—generalization that is sustainable. If a grammar is constructed by ranking a universal constraint set that includes only phonetically-motivated constraints (Fig. 5.2c), only the generalization from (125a) can be learned because that is the only phonetically-motivated generalization. On the other hand, if all relevant constraints are included together (Fig. 5.2d) or on their own (Fig. 5.2e), all three generalizations are learned roughly equally well. This is because learning reduces to constraint ranking—the constraints being provided—and thus fails to distinguish between more and less local constraints. For language 3, the number of exceptions against the more local generalization will eventually become too numerous under the Tolerance Principle, and PLP will construct a less-local rule, which will correctly characterize the language 3 alternation. Thus, PLP predicts that given sufficient time and data, learners will eventually be able to learn the language 3 alternation.

5.4.2.3 Takeaways

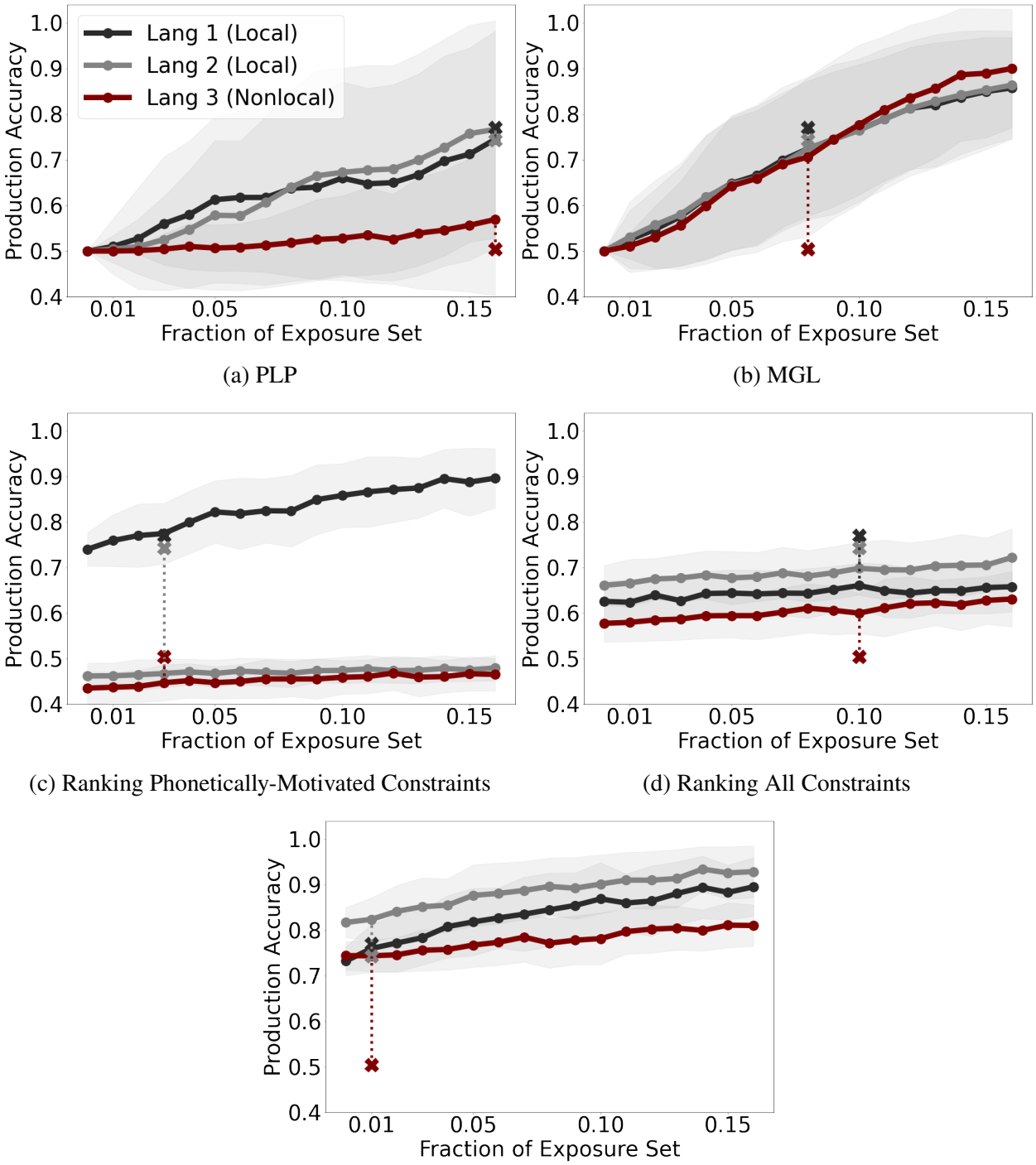
PLP reflects human learners' preference for local generalizations (**Q3**).

5.4.3 Learning German Devoicing

We now evaluate PLP on syllable-final obstruent devoicing in German (Wiese, 1996).

5.4.3.1 Setup

This experiment simulates child acquisition by using vocabulary and frequency estimations from child-directed speech in the Leo corpus (Behrens 2006). We retrieved the corpus from the CHILDES (MacWhinney 2000) database and intersected the extracted vocabulary with CELEX (Baayen et al. 1996) to get phonological and orthographic transcriptions for each word. We also computed the frequency of each word in the Leo corpus. The resulting dataset consists of 9,539 words. To construct URs and SRs, we followed Gildea and Jurafsky (1996), using the CELEX phonological representations as SRs and discrepancies between CELEX phonology and orthography to construct URs, since German orthography does not reflect devoicing. Specifically, we make



(a) PLP

(b) MGL

(c) Ranking Phonetically-Motivated Constraints

(d) Ranking All Constraints

(e) Ranking Language-Relevant Constraints

Figure 5.2: PLP best matches the locality results of Baer-Henney and van de Vijver (2012), where participants much more easily learned languages with local generalizations (languages 1-2) than a non-local generalization (language 3). In contrast, MGL fails to mirror these results, learning all generalizations equally well. Grammars constructed by ranking a provided constraint set also fail to match the results: if provided with phonetically-motivated constraints, only the first generalization can be learned, and if provided with all or language-relevant constraints, all generalizations are learned equally well.

the syllable-final obstruents voiced for the URs of all words where the corresponding orthography indicates a voiced obstruent. In this data, 8.2% of words involve devoicing, which means a substantial number of URs equal SRs w.r.t this process. However, this is an appropriate and realistic scenario, since the data was constructed from child-directed speech and is thus a reasonable approximation of the data that children have access to when learning this generalization.

The experimental procedure samples one word at a time from the data, weighted by frequency. The word is presented to each model and added to its vocabulary. Sampling is with replacement, so the learners are expected to encounter the same word multiple times, at frequencies approximating what a child would encounter. When the vocabulary reaches a size of 100, 200, 300, and 400, each model is probed to produce an SR for each UR in the dataset that is *not* in the vocabulary (i.e. held-out test data). The fraction of these predictions that it gets correct is reported as the model's accuracy. The models MGL, ED, and OSTIA are designed as batch learners, so they are trained from scratch on the vocabulary prior to each evaluation period.[16] PLP, UCON, and ORACLE learn incrementally.

This simulation is carried out 10 times to simulate multiple learning trajectories. The results are averages and standard deviations over these 10 runs.

ORACLE is provided with the constraint set (128).

(128) CON = { MAX, DEP, IDENT(VOICE), IDENT(SON), IDENT(NAS), *[+voi, −son]]$_\sigma$ }

The markedness constraint *[+voi, −son]]$_\sigma$, which marks syllable-final voiced obstruents, is the relevant markedness constraint for this process. UCON is provided two additional constraints: *NÇ, which marks voiceless consonants following nasals, and *COMPLEX. Both are frequently considered to be universal, violable constraints (Prince and Smolensky, 1993; Locke, 1983; Rosenthall, 1989; Pater, 1999). We included these to capture the assumption of a universal constraint set, which requires learning that *COMPLEX and *NÇ are violable in German; for instance /glaubənd/ → [glau.bənt.] ('believing') violates *COMPLEX and *NÇ.

5.4.3.2 Results

The results are shown in Tab. 5.1. PLP learns an accurate grammar, which consists of the single generalizaton shown in (129), where '$]_\sigma$' denotes the end of a syllable.

(129) [+voi,−son] → [−voi] / __ $]_\sigma$

While PLP achieves perfect accuracy by the time the vocabulary has grown to size 100, it does produce errors in the process of getting there. A primary example is underextensions. In our experiments, underlyingly voiced stops tended to enter the vocabulary earlier than voiced fricatives.

[16] We provide MGL the frequency with which each vocabulary word has appeared, which it can make use of.

Consequently, PLP sometimes fails to extend devoicing to fricatives until evidence of them devoicing enters the vocabulary. These underextensions are over *held-out test words*—i.e. words not in the learner's vocabulary. Thus, this is a prediction about an early state of the learner's phonological grammar, and not a prediction that children go through a stage of voicing final voiced fricatves. Indeed, as soon as an instance of fricative devoicing enters the vocabulary, we found that PLP extends the generalization to account for it.

Ranking a provided constraint set (ORACLE and UCON) can yield the same generalization as PLP: the sequence $[+\text{voi},-\text{son}]]_\sigma$ is not allowed in German and violations of this restriction are repaired by devoicing. But the differences in how PLP learns this generalization are informative. Both UCON and ORACLE are provided the knowledge that the sequence $[+\text{voi},-\text{son}]]_\sigma$ is marked. In contrast, PLP discovers the marked sequence in the process of learning.

In German, the onset [bl] is allowed (e.g., /blɑu/ → [blɑu]). PLP always produces the correct SR for /blɑu/ as a consequence of its identity default (Tab. 5.2). Whether a constraint-ranking model incorporates a preference for identity between inputs and outputs depends on what constraints it ranks. Because ORACLE ranks only the constraints active in the language being learned, it—like PLP—does not produce unmotivated errors. If a universal constraint set is ranked (UCON), then markedness constraints that are violable in the language being learned will lead to unmotivated errors. For instance, prior to downranking *COMPLEX, UCON sometimes produces [bəlɑu] for /blɑu/, with the complex onset /bl/ separated by a [ə], even though such onsets are allowed in German. However, deletion tends to be more common than epenthesis as a repair in child utterances, and it appears to occur due to articulatory limitations rather than by the child's hypothesized adult grammar.

Both UCON and ORACLE sometimes produce /kɪnd/ as *[kɪndə] and *[kɪn], rather than [kɪnt], because they must figure out the relative ranking of faithfulness constraints in order to capture which repair German uses to avoid $*[+\text{voi},-\text{son}]]_\sigma$ violations. In contrast, PLP infers the repair—devoicing—directly from what discrepancy it observes in the data.

None of the other models perform competitively: PLP outperforms them all by a statistically significant amount ($p < 0.01$), as measured by a paired t-test against the null hypothesis that each model's performance over the 10 simulations has the same average accuracy as PLP's. MGL, which generalizes as conservatively as possible, struggles to generalize beyond the training data. This is seen in its slow rate of improvement. ED is a powerful model in natural language processing when substantial amounts of data are available, but it struggles to learn on the small vocabularies at the scale children learn from. OSTIA struggles even more, consistent with the negative results of Gildea and Jurafsky (1996), who presented it with much larger vocabularies.

Table 5.1: Model accuracies (with standard deviations) on held-out test data at different training vocabulary sizes. PLP readily learns an accurate generalization for German final-obstruent devoicing.

Model	Vocabulary Size			
	100	200	300	400
PLP	**1.000 ± 0.00**	**1.000 ± 0.00**	**1.000 ± 0.00**	**1.000 ± 0.00**
MGL	0.918 ± 0.00	0.918 ± 0.00	0.918 ± 0.00	0.919 ± 0.00
ED	0.008 ± 0.00	0.178 ± 0.03	0.389 ± 0.04	0.543 ± 0.04
OSTIA	0.023 ± 0.02	0.022 ± 0.01	0.031 ± 0.01	0.040 ± 0.00
UCON	0.960 ± 0.03	0.988 ± 0.00	0.992 ± 0.00	0.995 ± 0.00
ORACLE	0.982 ± 0.01	0.997 ± 0.00	0.998 ± 0.00	0.999 ± 0.00
ID	0.918 ± 0.00	0.918 ± 0.00	0.918 ± 0.00	0.918 ± 0.00

Table 5.2: Analysis of the types of errors each of the models that learn an accurate grammar make in the process. Because it only adds generalizations to the grammar when necessitated by surface-alternation, PLP produces no unmotivated errors.

Error Type	Example	PLP	UCON	ORACLE
Unmotivated	/blɑu/ → *[bə̲lɑu]	No	Yes	No
Wrong-Repair	/kɪnd/ → *[kɪnd̲ə]	No	Yes	Yes
Under/Over Extension	/kɪnd/ → *[kɪn̲d]	Yes	Yes	Yes

5.4.3.3 Takeaways

PLP is readily able to learn German syllable-final devoicing (**Q2**) and never introduces unmotivated generalizations (**Q3**).

5.4.3.4 Opacity

Devoicing in Polish interacts opaquely with o-raising, in which /ɔ/ surfaces as [u] before final, underlyingly voiced, oral consonants (Kenstowicz, 1994; Sanders, 2003). As a proof-of-concept, we ran PLP on the data from Sanders (2003, chap 2; ex. 2-5). PLP learns rules (130) and correctly ordered them in COUNTERBLEEDING order with raising r_1 before devoicing r_2.[17]

[17]The examples from Sanders (2003) were too sparse to distinguish between [+voi]#and [+cons,+voi,−nas]#as the context for raising; a more realistic lexicon should drive PLP to the more nuanced context.

(130) $G = r_2 \circ r_1$, where

$$r_1 = \text{ɔ} \rightarrow \text{u} \;/\; \_\_ \; [+\text{voi}]\#$$
$$r_2 = [+\text{voi}, -\text{son}] \rightarrow [-\text{voi}] \;/\; \_\_ \; \#$$

Rule r_2 accounts for devoicing both in isolation (131a) and in words exhibiting raising (131c). Rule r_1 accounts for raising both in isolation (131b) and when its underlying context is opaquely obscured by devoicing (131c).

(131) a. /klub/ → [klup] 'club' SG
b. /bɔl/ → [bul] 'ache' NOM.SG
c. /bɔb/ → [bup] 'bean' NOM.SG

The correct ordering was achieved because, in the reverse ordering, devoicing bleeds raising, resulting in errors like *[bɔp] for /bɔb/ that are not present when in COUNTERBLEEDING order. This demonstrates that PLP is capable of handling at least this case of opacity. We leave a systematic study of opacity for future work (see § 5.2.3.3 and § 7.3.5).

5.4.4 Learning a Multi-Process Grammar

This experiment evaluates PLP at learning multiple generalization simultaneously. The processes modeled are the alternating plural and PRS 3RD SG affix /-z/ (132a), the alternating past tense affix /-d/ (132b), and vowel nasalization (132c).

(132) a. /dɑg-z/ → [dɑgz]
/wɔk-z/ → [wɔks]
/hɔrs-z/ → [hɔrsəz]
b. /smɛl-d/ → [smɛld]
/wɔk-d/ → [wɔkt]
/foʊld-d/ → [foʊldəd]
c. /ðɛm/ → [ðɛ̃m]
/sʌmθŋ/ → [sʌ̃mθĩŋ]
/dæns/ → [dæ̃ns]

5.4.4.1 Setup

This experiment, like the first, simulates child language acquisition. The child-directed speech is aggregated across English corpora in CHILDES (MacWhinney, 2000), including the frequency of

Table 5.3: Model accuracies (with standard deviations) on held-out test data at different training vocabulary sizes. PLP readily learns an accurate grammar for the English processes in (132).

Model	Vocabulary Size			
	1000	2000	3000	4000
PLP	**0.984 ± 0.01**	**0.992 ± 0.00**	**0.995 ± 0.00**	**0.997 ± 0.00**
UCON	0.969 ± 0.00	0.982 ± 0.00	0.987 ± 0.00	0.990 ± 0.00
ORACLE	0.980 ± 0.00	0.989 ± 0.00	0.991 ± 0.00	0.992 ± 0.00
ID	0.510 ± 0.00	0.510 ± 0.00	0.510 ± 0.00	0.510 ± 0.00

each word. Only words with '%mor' tags were retained, because the morphological information was needed to construct URs. Transcriptions from the CMU pronunciation dictionary (CMU, 2014) served as SRs, with nasalization added to vowels preceding nasal consonants. URs had all vowels recorded without nasalization. The surface affixes for all past tense verbs, plural nouns, and PRS 3RD SG verbs were set to /d/, /z/, and /z/, respectively in the URs. The resulting dataset contains 20,421 UR-SR pairs.

The experimental procedure is the same as for German, sampling words weighted by frequency and reporting accuracies at predicting SRs from URs over held-out test words when each learner's vocabulary reaches certain sizes: 1K, 2K, 3K, and 4K words.

We omit results from MGL, ED, and OSTIA because they continued to be noncompetitive. ORACLE once again ranks only the relevant constraints (133) and UCON receives, in addition to (133), *COMPLEX and *NÇ.

(133) CON = {
MAX, DEP, IDENT(VOICE), IDENT(SON), IDENT(NAS),
AGREE(VOICE), *SS, *[+vowel, −nas][+cons, +nas],
*[−cont,−dist,−son][−cont,−dist,−son]
}

All faithfulness constraints other than DEP were split into two—one for stems and one for affixes—so that, for instance, *[wɔgz] could be ruled out for input /wɔk-z/ in (132).

5.4.4.2 Results

The models' accuracies on held-out test words, shown in Tab. 5.3, reveal that PLP learns an accurate grammar by the time its vocabulary grows to about 2000 words. PLP's output is shown as an ordered list of rules in (134).

(134) $G = r_5 \circ r_4 \circ r_3 \circ r_2 \circ r_1$, where

$$r_1 = [+\text{syl}] \rightarrow [+\text{nas}] \,/\, \_\_\, [+\text{nas}]$$
$$r_2 = \emptyset \rightarrow \text{ə} \,/\, [+\text{sib}]\, \_\_\, [+\text{sib}]$$
$$r_3 = [+\text{sib}, +\text{voi}] \rightarrow [-\text{voi}] \,/\, [-\text{voi}]\, \_\_$$
$$r_4 = \emptyset \rightarrow \text{ə} \,/\, [+\text{cor}, -\text{cont}, -\text{nas}]\, \_\_\, [+\text{cor}, -\text{cont}, -\text{nas}]$$
$$r_5 = [+\text{cor}, -\text{cont}, -\text{nas}, +\text{voi}] \rightarrow [-\text{voi}] \,/\, [-\text{voi}]\, \_\_$$

The rules were ordered as described in § 5.2.3.3, with r_2 before r_3 and r_4 before r_5 (i.e. BLEEDING order) being the inferred ordering dependencies. Thus, as described in § 5.2.3.3, PLP learned that epenthesis bleeds devoicing. The rules r_2-r_5 do not encode a word-final context because doing so would require expanding PLP's search window, which is not necessary because the rules without word-final context pass the Tolerance Principle. The extension of [+cor,−cont,−nas] is {t, d} and of [+cor,−cont,−nas,+voi] is {d}.

The reason no model achieves 100% accuracy is due to a handful of words that do not follow the generalizations in (132). For instance, compounds like [bɛdtaĩm] allow the sequence [dt], but the models predict there should be an epenthetic vowel to split the sequence. Such exceptions are easily accounted for if we assume the learner recognizes the word as a compound. Since exceptions are inevitable in naturalistic data, we chose to not remove these exceptions.

In Berko (1958)'s seminal study, Berko found that children aged 4-7yrs could accurately inflect nonce words that take the [-z], [-s], [-d], or [-t] suffixes, but that they performed much worse at inflecting nonce words taking the [-əz] or [-əd] suffixes. Adults could inflect nonce words with [-əz] or [-əd], suggesting that voicing assimilation process may be learned earlier than the epenethesis process. We show PLP's accuracy on Berko (1958)'s different categories of nonce words in Fig. 5.3 as the vocabulary grows (x-axis). PLP's accuracy on nonce words taking [-z] or [-s] (black dashed line) converges earlier than its accuracy on nonce words taking [-əz] (gray dashed line); similarly the accuracy for nonce words taking [-d] or [-t] (black dotted line) converges earlier than for nonce words taking [-əd] (gray dotted line). Thus, the order of acquisition matches Berko (1958)'s finding.

5.4.4.3 Takeaways

The results in this more challenging setting, where multiple processes are simultaneously active, support the takeaways from the prior experiment. PLP successfully learns all the generalizations (**Q2**) and does not introduce unmotivated generalizations (**Q3**).

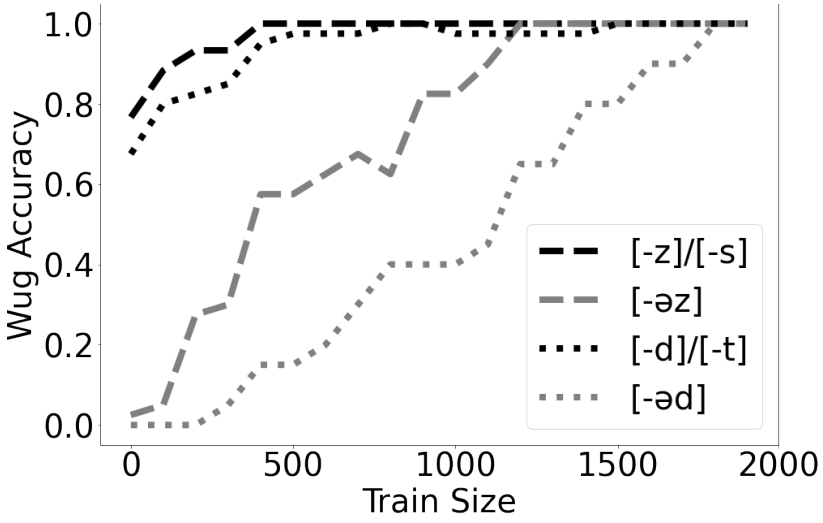


Figure 5.3: PLP's accuracy on the plural and past tense nonce words from Berko (1958) as training progressed. The black dashed line denotes plurals that should take [-z] or [-s] and the gray dashed lines those that should take [-əz]. The dotted lines represent the analogues for past-tense. The fact that [z]/[s] accuracy converges before [-əz] and [d]/[t] before [-əd] matches Berko (1958)'s finding that children learn [-z]/[-s] and [-d]/[-t] before [-əz] and [-əd].

5.4.5 Learning Tswana's Post-Nasal Devoicing

Although a majority of phonological patterns may be phonetically grounded, some processes nevertheless appear to lack or even oppose phonetic motivation (Anderson, 1981; Buckley, 2000; Johnsen, 2012; Beguš, 2019). Moreover, these must still be learnable, because children continue to successfully acquire them (Johnsen, 2012, p. 506). An example of such a pattern is post-nasal devoicing in Tswana shown in (90), which Coetzee and Pretorius (2010) confirmed to be productive despite operating against the phonetic motivation for post-nasal *voicing* (Hayes and Stivers, 2000; Beguš, 2019). Beguš (2019, p. 699) found post-nasal devoicing to be reported as a sound change in thirteen languages and dialects, from eight language families.

Models of phonological learning should account for the fact that non-phonetically-grounded, yet productive patterns are successfully learned by humans. A consequence of PLP's identity default is that generalizations are added to the grammar whenever they are motivated by surface alternation. Since surface alternation in Tswana motivates a generalization for post-nasal devoicing, its learnability should be accounted for with PLP. This experiment attempts to confirm this (**Q3**).

5.4.5.1 Setup

For this experiment we used the 10 UR-SR pairs from Coetzee and Pretorius (2010, p. 406) as training data. Five pairs involve devoicing resulting from the 1ST SG OBJ clitic /m/ attaching to a stem

Table 5.4: PLP learns precisely the set of processes active in its experience. This provides a straightforward account of how productive phonological processes can be learned even if they operate against apparent phonetic motivation, like devoicing in Tswana following nasals (Coetzee and Pretorius, 2010). With PLP, the unmotivated constraint *NC̬ need not be assumed universal.

Model	Generalization	Test Accuracy
PLP	b → [−voi] / m __	1.0
Ranking without *NC̬	{*NC̬, IDENT(VOICE)}	0.5
Ranking with *NC̬	*NC̬ » {*NC̬, IDENT(VOICE)}	1.0

that starts with a voiced obstruent. The other five pairs involve the 1ST PL OBJ clitic /re/ attaching to the same stems, which serve as negative examples since the clitic does not introduce a nasal. This data is not necessarily representative of the data that a child would have during acquisition, and thus serves as a proof-of-concept learnability experiment.

The test data consists of the same 20 /b/-initial nonce words presented to the participants in Coetzee and Pretorius (2010, p. 407)—10 stems each combined with /m/ and /re/.

5.4.5.2 Results

The results in Tab. 5.4 demonstrate that PLP can learn Tswana's post-nasal devoicing without requiring the existence of a universal, phonetically unmotivated constraint.[18] Constraint-ranking models can also learn the generalization, but depend on an account of how the constraint *NC̬, which is not usually considered to be a universally marked sequence (Locke, 1983; Rosenthall, 1989; Pater, 1999; Beguš, 2016, 2019), is added to the constraint set.

5.4.5.3 Takeaways

Because PLP assumes UR-SR identity by default, it constructs precisely the generalizations necessary to account for the discrepancies active in its experience, providing a straight-forward account of how productive generalizations can be learned even if they are opposed to apparent phonetic motivation, as humans evidently do (Seidl and Buckley 2005, Johnsen 2012, p. 506; Beguš 2018, ch. 6) (**Q3**).

[18]PLP learns *[mb] rather than *NC̬ because the training data only included [mb] instances; if more representative training data were available, PLP would induce natural classes, as in the previous experiments.

5.5 Discussion

One reviewer of our Belth (2023a) paper asked what sort of tendency we view locality to be. We view the cognitive tendency for humans to prefer constructing local generalizations to be a *gemoetric, computational* consequence. That is, if words are viewed, at least to a first-approximation, as linear objects, this linear geometry introduces the notion of locality as *small linear distance.* In our view, the reason that a human is more likely to construct a generalization that conditions x_i on x_{i-1} than on x_{i-2} in a sequence $..., x_{i-2}, x_{i-1}, x_i$ (see § 5.1.1) is that a search outward from x_i encounters x_{i-1} before it encounters x_{i-2}. PLP is an attempt to state this in explicit computational terms. An immediate consequence of this hypothesis is that if x_{i-1} is sufficient to account for whatever the uncertainty in x_i is (e.g., what its surface form is), then x_{i-2} will never be considered, even if there is some statistical dependency between the two. We believe this prediction is consistent with the experimental results from sequence learning, which we discuss in § 1.1, where participants would track adjacent dependencies even when non-adjacent dependencies were more statistically informative (Gómez and Maye, 2005) and would construct local generalizations over less local ones when the exposure data underdetermined the two (i.e. the poverty-of-stimulus paradigms of Finley 2011, McMullin and Hansson 2019). We further confirm this in chapter § 6. We note that words may not be *exactly* linear—segment articulations have gestural overlap, syllables are often viewed as hierarchical structures, and representations like autosegmental tiers may be present. However, we think treating words as linear sequences is a good first approximation. Work on tier-locality also recognizes that string-locality is a special case of tier locality in which all segments are present on the tier (e.g. Hayes and Wilson 2008; Heinz et al. 2011; McMullin 2016).

An alternative view could be that locality is distributional: a learner may track the dependency between x_i and both x_{i-1} and x_{i-2}, and may find that x_{i-1} is more statistically robust as a generalization, preferring it for that reason. However, this view is inconsistent with the findings that when statistical robustness is controlled (Finley, 2011; McMullin and Hansson, 2019) and even when it *favors* the less-local dependency (Gómez and Maye, 2005) humans systematically generalize locally. The distributional approach could be combined with a stipulated bias (prior) favoring local dependencies, but this would simply describe the phenomena, not explain it.

5.5.1 Future Directions

Research on phonological representations recognizes that strict locality arises not only over string representations, but also over representations like tiers and metrical grids (Goldsmith, 1976; Heinz et al., 2011; Hayes and Wilson, 2008; McMullin, 2016). PLP could be applied over these representations to find, e.g. tier-strictly local generalizations. In chapter § 3, we provided a unified model of local and non-local learning which captures precisely this idea. We discuss how these pieces may

fit together in § 7.3.1.

CHAPTER 6

Experimental Evaluation

The models in § 3 and § 5 track dependencies adjacent to an alternating segment first, and only expand the search outward (§ 5) or change representations (§ 3) when necessary. This procedure predicts that if adjacent dependencies are sufficient to predict an alternation, the learner will not discover possible dependencies at longer distances, even if they are as statistically robust. In this chapter, we design and perform an artificial language experiment to test precisely this prediction. The artificial language exhibits an alternation in the formation of plural nouns via the suffixation of one of two phonologically-conditioned suffixes to a noun stem. The exposure data presented to participants underdetermines the generalization underlying plural formation, such that both the final segment and the penultimate segment of the stem equally predict the form of the plural suffix. The test data evaluates which generalization learners construct. Our models' predictions are borne out, as participants show evidence of generalizing based on the stem-final segment, even though the penultimate segment provided equally statistically-robust information. We discuss further implications of the results, including how they relate to proposals that learners generalize as conservatively as possible, how human generalization differs from the behavior of transformer and n-gram language models, and how our algorithmic approach to phonology complements formal-language-theoretic characterizations of phonology.

6.1 Introduction

If we consider again a local alternation like the English plural, repeated in (135), it is possible that learners track dependencies between every stem segment and the plural suffix, and then learn to attend specifically to the stem-final segment because of its comparative statistical robustness at predicting the suffix's realization.

(135) [dɑg̲z]
[kæt̲s]
[hɔrs̲əz]

Our models in § 3 and § 5 track adjacent dependencies first, and thus predict that this is not the case. In this chapter, we develop an artificial language experiment that creates a scenario allowing these alternatives to be decisively evaluated. Consider the schema of an alternation where a becomes b when preceded by c, but where c is *always* preceded by a d.[1]. Descriptively, in rule notation, this could be characterized equally well as any of the rules in (136), where [$*$] stands for any segment.

(136) a. $a \rightarrow b \ / \ c\_\_$

b. $a \rightarrow b \ / \ dc\_\_$

c. $a \rightarrow b \ / \ d[*]\_\_$

The iterative nature of PLP and D2L, which both start by tracking adjacent dependencies, will form rule (136a), because that is sufficient to account for the pattern in the exposure data. We will call this generalization LOCAL.

Many linguists, in phonology (e.g., Albright and Hayes 2003; Hale and Reiss 2008) and syntax (e.g., Berwick 1985), have argued that learners generalize as *conservatively* as possible, often because of a theoretical learning problem called the *Subset Principle*, which arises when learning from positive data only. Generalizing as conservatively as possible leads to rule (136b), which avoids generalizing to novel instances where c is not preceded by a d.

The generalization (136c) would be a surprising result of this scenario, but it could correspond to a scenario where a learner constructs a tier that includes the segments in d and in b while excluding those in c. We will thus call this generalization TIER.

If the learner proceeds in the fashion we suggested at the beginning of this section—tracking all dependencies and homing in on the most statistically robust, then this suggests that there should be no preference among the generalizations, because they are all equally robust. We will call this hypothesis STATROBUST.

For novel words in which a is preceded by c, but c is *not* preceded by d—e.g. xca for some $x \neq d$—LOCAL makes a prediction distinct from the other three generalization strategies, as summarized in (137).

(137) LOCAL: $xca \rightarrow xcb$, $*xca$

CONSERVATIVE: $xca \rightarrow xca$, $*xcb$

TIER: $xca \rightarrow xca$, $*xcb$

STATROBUST: No preference between xca and xcb

This chapter presents an experimental study to test precisely this prediction, thus serving as an evaluation of these alternative generalization strategies, and an attempt to falsify our models'

[1]Here a, b, c, and d are just stand-ins for some set of segments.

prediction. Because the training data underdetermines the generalization, the experiment will use the 'poverty of stimulus' artificial language paradigm to evaluate the alternative predictions. This paradigm has been used extensively for evaluating what generalizations learners form when multiple are consistent with exposure data (Finley, 2011, 2015, 2017; McMullin and Hansson, 2019).

In § 6.2, we provide details of how the experiment was conducted, including a description of the artificial language matching the (136) schema (§ 6.2.1). In § 6.3 we present the results. We then compare PLP from chapter § 5, along with n-gram and transformer language models, to the human behavior. We conclude in § 6.5 with a discussion of the results, and their implications.

6.2 Methodology

6.2.1 The Artificial Language

The artificial language used in the study forms plurals by adding a suffix[2], which alternates between [-f] and [-ʃ]. In the training data, exemplified in (138), [-f] surfaces as the PL affix whenever the stem (singular) ends in a voiced consonant and a back vowel (138a); [-ʃ] surfaces whenever the stem ends in a voiceless consonant and a front vowel (138b).

(138) a. [bibu] ~ [bibuf]
[bɔtbɔ] ~ [bɔtbɔf]

b. [pɔti] ~ [pɔtiʃ]
[dubtɛ] ~ [dubtɛʃ]

This matches the schema of (136). The LOCAL generalization strategy, predicted by PLP and D2L, will determine the form of the affix in terms of the final vowel because that is the closest element to the alternating affix. In contrast, the CONSERVATIVE generalization strategy will take the intersection of the shared environments, and determine the form of the affix in terms of both the final segment (vowel) and the penultimate segment (consonant). The TIER generalization strategy will determine the form of the affix in terms of the penultimate consonant of the stem, effectively creating a consonant tier to skip over the final vowel. The STATROBUST strategy will determine the form of the suffix based on any of these equally-statistically-robust characteristics of the stem.

To make this more concrete, we will suppose that /-ʃ/ is interpreted as the default (*elsewhere*) suffix by the learners. This is plausible, since [ʃ] tends to have higher amplitude and longer duration than [f] (Behrens and Blumstein, 1988) and is closer to the English-plural [-s]. We assume this for ease of exposition, but consider alternatives in our analysis (§ 6.3).

[2]The choice of the affix's morphological purpose as marking plurals is clearly arbitrary.

Thus, the LOCAL generalization is shown in rule-notation in (139a), along with the CONSERVATIVE generalization in (139a) and the TIER generalization in (139c).

(139) a. LOCAL
/ʃ/ → [f] / [+vowel,+back] __
Elsewhere [ʃ]

b. CONSERVATIVE
/ʃ/ → [f] / [+cons,+voi][+vowel,+back] __
Elsewhere [ʃ]

c. TIER
/ʃ/ → [f] / [+cons,+voi][*] __
Elsewhere [ʃ]

All generalizations are consistent with the training data. However, the LOCAL generalization make different predictions about participant behavior on a novel test items where a back vowel is preceded by a voiceless consonant ([+cons,−voi]) instead of a voiced consonant ([+cons,+voi]). The LOCAL generalization predicts that [-f] will be the surface form of the PL suffix when the final vowel is back, even if the penultimate consonant is voiceless (140a). In contrast, the CONSERVATIVE generalization will not apply to such test items, and thus predicts that default (elsewhere) [-ʃ] will be the PL suffix (140b). Similarly, because the TIER generalization tracks the dependency between the penultimate consonant and the suffix, it too predicts that these items, which do not match its structural description, will take the default [-ʃ] form (140c).

(140) a. LOCAL
/dupu-ʃ/ → [dupuf]
/pɔdpu-ʃ/ → [pɔdpuf]

b. CONSERVATIVE
/dupu-ʃ/ → [dupuʃ]
/pɔdpu-ʃ/ → [pɔdpuʃ]

c. TIER
/dupu-ʃ/ → [dupuʃ]
/pɔdpu-ʃ/ → [pɔdpuʃ]

The choice of [-f]/[-ʃ] was chosen because it is phonetically unmotivated: there is no clear reason for these two segments to alternate, and there is no clear relationship between the alternation and the segments in the environment that determine the alternation. We chose an unmotivated alternation so that if learners indeed go with the LOCAL generalization, it will likely be because of the locality of the generalization, not the fact that dependencies between adjacent segments are phonetically

Table 6.1: Segment Inventory

Natural Class	Description	Extension
C	consonants	{t, d, p, b}
[+cons,+voi]	voiced consonants	{d, b}
[+cons,−voi]	voiceless consonants	{t, p}
V	vowels	{i, ɛ, u, ɔ}
[+vowel,+back]	back vowels	{u, ɔ}
[+vowel,−back]	front vowels	{i, ɛ}

Table 6.2: Training Data

Stem	Suffix	Number	Example
CV.[+cons,+voi][+vowel,+back]	[-f]	25	(bibu, bibuf)
CVC.[+cons,+voi][+vowel,+back]	[-f]	25	(bɔtbɔ, bɔtbɔf)
CV.[+cons,−voi][+vowel,−back]	[-ʃ]	25	(pɔti, pɔtiʃ)
CVC.[+cons,−voi][+vowel,−back]	[-ʃ]	25	(dubtɛ, dubtɛʃ)

motivated. This choice also allows for further validation of the the results of other experiments by Seidl and Buckley (2005) and Beguš (2018), which demonstrate that humans can learn phonological processes lacking phonetic motivation.

6.2.2 Participants

Participants were recruited on Prolific and compensated at a rate of $16 per hour. The participants were adults ages 18-50 who were L1 speakers of English. They were required to use a desktop equipped with audio during the task.

6.2.3 Stimuli

Stimuli were formed from the segment inventory described in Tab. 6.1. The training stems, summarized in Tab. 6.2, either end in a voiced consonant and a back vowel or a voiceless consonant and a front vowel. The former take a [-f] PL affix and the later a [-ʃ] PL affix. Consequently, the training data is consistent with any of the generalization strategies described in the preceding sections.

The training data consists of 100 <stem, stem+PL > pairs; 50 of the stems end in [+cons,+voi][+vowel,+back] and 50 end in [+cons,−voi][+vowel,−back]. To further increase the variety of forms, each of these groups has 25 stems that start with a CV syllable and 25 that start with a CVC syllable.

Table 6.3: Test Data

	Stem	Choices	Num	Example
Novel	CV.[+cons,−voi][+vowel,+back]	[-f]/ [-ʃ]	10	(dupu, dupuf/dupuʃ)
	CVC.[+cons,−voi][+vowel,+back]	[-f]/ [-ʃ]	10	(pɔdpu, pɔdpuf/pɔdpuʃ)
	CV.[+cons,+voi][+vowel,−back]	[-f]/ [-ʃ]	10	(tidi, tidif/tidiʃ)
	CVC.[+cons,+voi][+vowel,−back]	[-f]/ [-ʃ]	10	(pɛpdɛ, pɛpdɛf/pɛpdɛʃ)
Train-Like	CV.[+cons,+voi][+vowel,+back]	[-f]/ [-ʃ]	3	(pɛbɔ, pɛbɔf/pɛbɔʃ)
	CVC.[+cons,+voi][+vowel,+back]	[-f]/ [-ʃ]	3	(ditbu, ditbuf/ditbuʃ)
	CV.[+cons,−voi][+vowel,−back]	[-f]/ [-ʃ]	3	(putɛ, putɛf/putɛʃ)
	CVC.[+cons,−voi][+vowel,−back]	[-f]/ [-ʃ]	3	(tɛpti, tɛptif/tɛptiʃ)

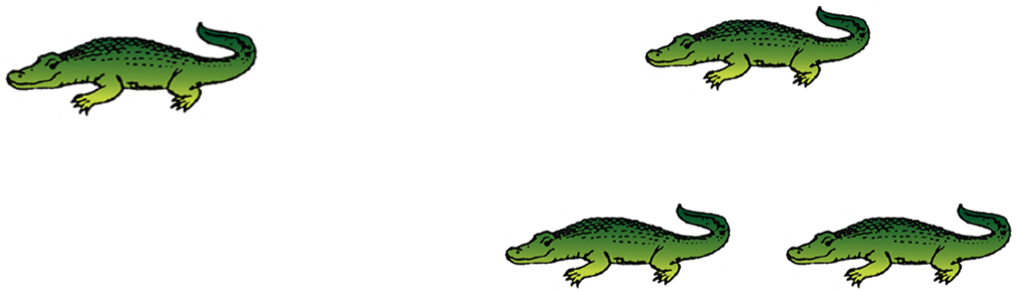
The test data—described in Tab. 6.3—consists of 52 stems, each with two possible PL forms ending in [-f]/[-ʃ]. Of these, 40 are novel items where—unlike the training data—back vowels are proceeded by voiceless consonants and front vowels by voiced consonants. For these forms, LOCAL makes predictions that differ from those of the other possible generalizations, as described in § 6.3. Of these, 20 end in [+cons,−voi][+vowel,+back] and 20 in [+cons,+voi][+vowel,−back]; within each group of 20, 10 start with a CV syllable and 10 with a CVC syllable. The remaining 12 items are 3 forms from each of the four categories of training instances. These forms have the same structure as training items, but are new words, thus requiring generalization.

To avoid English words entering the data, words where the stem or either of its possible affixed forms existed in the CMU (2014) pronunciation dictionary were not allowed. The full training and test materials are listed in Tables C.1-C.2 in the appendix. The stimuli were recorded in the sound lab at the University of Michigan Linguistics Department.

Each word was paired with an image of a common object—one item for singular nouns, and three of the same item for corresponding plural nouns. Following Baer-Henney and van de Vijver (2012), the images were created from the color versions of the Snodgrass and Vanderwart (1980) collection created by Rossion and Pourtois (2004) and available through the University of Lorraine. The images were pre-filtered to remove images where the singular item could be ambiguous as to its number (e.g., a singular hand could be misconstrued as multiple fingers). An example is shown in Fig. 6.1, where the plural word [pidpiʃ] contains three of the items (alligators) in the singular [pidpi] image. The images were paired with the words at random.

6.2.4 Experimental Design

Below we discuss the experiment design, which includes a training phase followed by a testing phase.



(a) Image for singluar noun [pidpi].

(b) Image for plural noun [pidpiʃ].

Figure 6.1: Example images for stimuli.

6.2.4.1 Training Phase

The first phase involved presenting nouns to participants. Participants in the experimental group were presented the 100 training nouns summarized in Tab. 6.2. Each singular noun was followed by its plural form, separated by a 500ms pause. Each noun was accompanied by the picture capturing the word's meaning, which switched from the singular image to the plural image at the onset of the plural stimulus.

Participants in the control group were presented the same singular nouns (including images) as the experimental group, but no plural forms. The training data were presented to each participant in a random order, reshuffled for each participant.

The training phase was self-paced, with participants pressing the spacebar between each noun (singular-plural pair for the experimental group, or singular noun for the control group) to continue.

6.2.4.2 Testing Phase

After training, both the experimental and control groups were tested on the same test items—those summarized in 6.3—in a sequence of two-alternative forced choice (2AFC) tests. Each singular noun was followed by two options for the PL form, one ending in [-f] and the other ending in [-ʃ]. The first option followed the singular after a 500ms pause, and a second 500ms pause separated the second option from the first. The order in which the [-f] and [-ʃ] choices were presented was randomized. After the presentation of the second choice, the participants had to select which choice they thought was the plural form of the noun in the language they had just learned. Their choice was entered by pressing 'a' on their keyboard for the first choice or 'b' for the second choice. After the participant made their selection, a 500ms pause preceded beginning of the next trial.

The train-like test items, which are of the same form as the training items, test whether partic-

ipants generalized from the training data. The novel test items test which generalization strategy participants used.

6.2.5 Hypotheses

LOCAL, CONSERVATIVE, TIER, and STATROBUST make distinct predictions about how learners will generalize from the training data to the test data. Here we outline the pattern in participant responses predicted by each of these hypotheses. The pattern of responses depends on whether participants treat one of the affixes as the default and, if so, which one. We describe three possible scenarios in (6.2.5.1)-(6.2.5.3), and the predicted affix form for test items is summarized in Tab. 6.4 for each hypothesis and scenario.

6.2.5.1 Scenario 1: /-ʃ/ treated as default (underlying) affix

The generalizations predicted by the LOCAL, CONSERVATIVE, and TIER hypotheses for this scenario were described in (139), as repeated in (141). The STATROBUST predicts that learners will show no preference for any of these over the others.

(141) a. LOCAL
/ʃ/ → [f] / [+vowel,+back] __
Elsewhere [ʃ]

b. CONSERVATIVE
/ʃ/ → [f] / [+cons,+voi][+vowel,+back] __
Elsewhere [ʃ]

c. TIER
/ʃ/ → [f] / [+cons,+voi][*] __
Elsewhere [ʃ]

Novel [+cons,+voi][+vowel,−back]-final test items fall under the elsewhere condition of both LOCAL and CONSERVATIVE, so both of these hypotheses predict the default [-ʃ] form for such trials. These test items match the structural description of TIER, which predicts the [-f] form for these items. On the other hand, novel [+cons,−voi][+vowel,+back]-final test items fall under the rule for LOCAL but the elsewhere condition for both CONSERVATIVE and TIER. Thus, CONSERVATIVE and TIER predict that these items will take default [-ʃ], while LOCAL predicts that the default /-ʃ/ will be realized as [-f] for such items. STATROBUST predicts no consistent preference for either type of novel test item.

6.2.5.2 Scenario 2: /-f/ treated as default (underlying) affix

This scenario is symmetrical with the last, with the role of [-f] and [-ʃ] reversed, as shown in (142).

(142) a. LOCAL
/f/ → [ʃ] / [+vowel,+back] __
Elsewhere [f]

b. CONSERVATIVE
/f/ → [ʃ] / [+cons,+voi][+vowel,+back] __
Elsewhere [f]

c. TIER
/f → [ʃ] / [+cons,+voi][*] __
Elsewhere [f]

Both LOCAL and CONSERVATIVE predict [-f] for novel [+cons,+voi][+vowel,−back]-final test items, which fall under their elsewhere conditions, while TIER predicts [-ʃ] for these items. For novel, [+cons,−voi][+vowel,+back]-final test items, CONSERVATIVE predicts that these forms will take default [-f], while CONSERVATIVE and TIER predict that they will take [-ʃ]. Again, STATROBUST predicts no consistent preference for either.

6.2.5.3 Scenario 3: Neither affix treated as default

If neither form is treated as the default, then the learners would be learning the conditioning of each affix as separate generalizations, as shown in (143).

(143) a. LOCAL
[-f] / [+vowel,+back] __
[-ʃ] / [+vowel,−back] __

b. CONSERVATIVE
[-f] / [+cons,+voi][+vowel,+back] __
[-ʃ] / [+cons,−voi][+vowel,−back] __

c. TIER
[-f] / [+cons,+voi][*] __
[-ʃ] / [+cons,−voi][*] __

Novel, [+cons,−voi][+vowel,+back]-final test items fall under the first generalization of LOCAL but the second of TIER, while [+cons,+voi][+vowel,−back]-final test items fall under the second generalization of LOCAL and the first of TIER. Neither of the CONSERVATIVE generalizations apply to either type of novel test item, since the voicing of the consonants preceding back/front vowels is

Table 6.4: The affix predicted by LOCAL and CONSERVATIVE for Experimental Group participants when presented with Novel Test instances. Cells with '??' denote a prediction that the Experimental Group should show no preference for one affix over the other.

Default	Generalization	Novel Test Item Ending	
		[+cons,−voi][+vowel,+back]	[+cons,+voi][+vowel,−back]
/-ʃ/	LOCAL	[-f]	[-ʃ]
	CONSERVATIVE	[-ʃ]	[-ʃ]
	TIER	[-ʃ]	[-f]
	STATROBUST	??	??
/-f/	LOCAL	[-f]	[-ʃ]
	CONSERVATIVE	[-f]	[-f]
	TIER	[-ʃ]	[-f]
	STATROBUST	??	??
None	LOCAL	[-f]	[-ʃ]
	CONSERVATIVE	??	??
	TIER	[-ʃ]	[-f]
	STATROBUST	??	??

flipped from that in the training data. In such a scenario, then, participants would have to resort to guessing the plural form of the novel items. Thus, both CONSERVATIVE and STATROBUST predict no consistent preference for either type of novel test item.

6.2.5.4 Summary

The affixes that each hypothesis predicts the Experimental Group will choose for Novel Test items is summarized in Tab. 6.4. Across all scenarios, LOCAL always predicts that Experimental Group participants will generalize to the Novel Test items based on the affix-adjacent vowel, while TIER predicts that they will generalize based on the penultimate consonant of the stem, and CONSERVATIVE predicts that they will apply the default affix (if there is one) to Novel Test items, or show no systematic preference for one affix over the other if participants learn no default affix. Since any of these generalizations are consistent with the training data, if learners track the most statistically-robust dependencies, as captured by the STATROBUST hypothesis, we would expect the participants to show no preference for either form. Critically, LOCAL's predictions—which are PLP and D2L's—are distinct from all other generalization strategies, regardless of which of the three possible scenarios (§ 6.2.5.1-6.2.5.3) we are in.

Aligning these predictions with our analysis of responses requires identifying which scenario manifested; evaluating the predictions statistically requires knowing, for reference, whether the

control group systematically preferred one affix or neither.

To evaluate the hypotheses, our analysis thus follows three steps. First, we identify whether the experimental group and control group treated either affix as default and, if so, which. Second, we confirm that training is effective for the experimental group, allowing them to construct a generalization consistent with the training data. These prerequisite steps established which of the above three scenarios we are in, and what the baseline behavior of the control group is. When this information is in hand, we will re-state each of the hypotheses in explicit terms, and evaluate which is supported by the results.

6.3 Evaluation

We first evaluate whether participants learned a generalization from the training data, and whether learners treated either suffix as the default (§ 6.3.1). We then evaluate whether the results support the LOCAL hypothesis (§ 6.3.2).

We gathered responses from 60 participants who were randomly assigned to either the control or experimental group. We ended up with 27 participants in the experimental group and 33 in the control group. The average age of participants was 31.5 years old; 34 participants reported Male sex, and 26 reported Female. Two of the participants who listed Male as their sex reported identifying as non-binary gender.

6.3.1 Effective Learning and Default Affix

In Fig. 6.2, we plot the distribution of participants' accuracies at choosing the training-like test form consistent with the training data. That is, the fraction of training-like test items that the participant chose the [-f] form if the stem ended in [+cons,+voi][+vowel,+back] or the [-ʃ] form if the stem ended in [+cons,−voi][+vowel,−back]. Most experimental-group participants achieve over 0.5 accuracy, with many achieving 1.0 accuracy. In contrast, control-group performance appears centered around chance performance, as would be expected by random guessing or systematically choosing based on a chosen default (e.g., [-ʃ]). This strongly suggests that the training phase effectively leads many experimental-group participants to construct a generalization consistent with the training data.

We show, in Fig. 6.3, the distribution of which form ([-ʃ] or [-f]) was chosen in the 2AFC tests. The control group, who received no information about plural formation during the training phase, systematically preferred the [-ʃ] form, as can be seen in Fig. 6.3a and its complement Fig. 6.3b. The experimental group also appears to have some preference for the [-ʃ] form, but this effect is much smaller than the control group. These distributions suggest that participants in both groups treat [-ʃ]

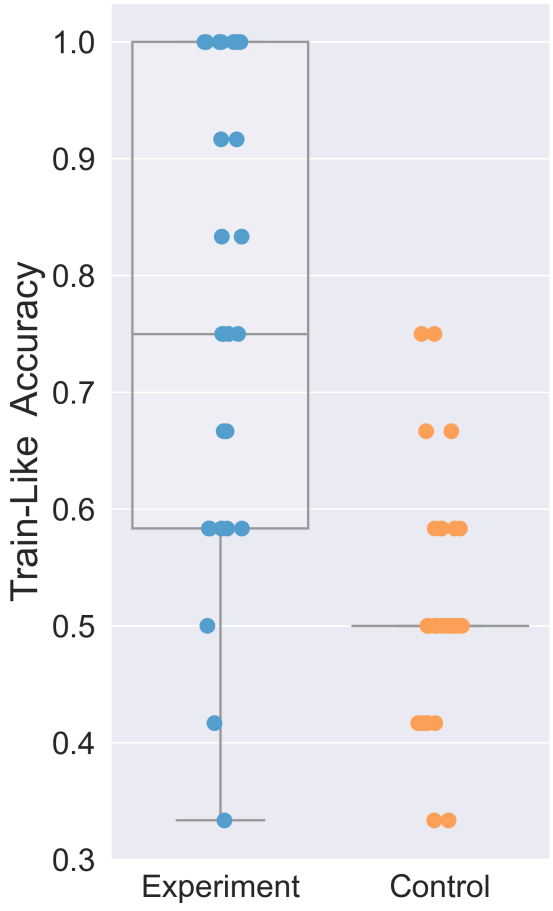
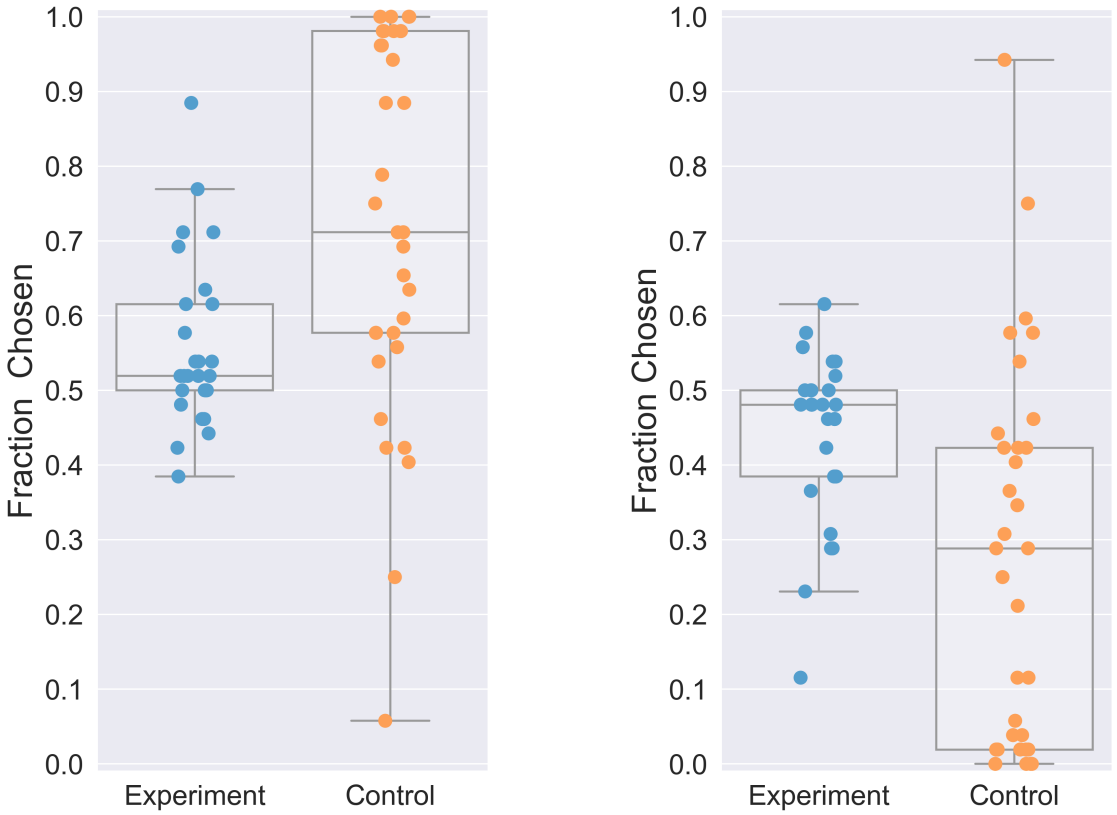


Figure 6.2: Accuracy on Train-Like Test Items.



(a) Fraction of responses where [-ʃ] form chosen

(b) Fraction of responses where [-f] form chosen

Figure 6.3: The distributions of which affix form was chosen.

Table 6.5: Fixed-effects component of the mixed model fit to the responses to train-like test items.

	Estimate	Std. Err	Pr(> |z|)
Intercept	0.05215	0.14801	0.72500
Exp	1.32369	0.23568	< 0.00001***
[+cons,+voi][+vowel,+back]	−0.97765	0.11588	< 0.00001***
Exp × [+cons,+voi][+vowel,+back]	0.70843	0.18065	0.00009***

as the default, and, since the preference for [-ʃ] is much smaller in the experimental group, support the conclusion that the training stage is effective.

Participants treating [-ʃ] as the default also makes sense given that [ʃ] tends to have higher amplitude and longer duration than [f] (Behrens and Blumstein, 1988), which could increase its saliency, and, as a sibilant, [-ʃ] is closer to the English plural [-s].

To analyze this results statistically, we used the *glmer* function of the *lme4* package (Bates et al., 2014) in R to fit a mixed-effects logistic regression model to the responses for train-like test items. The categorical, response variable was whether, on a particular 2AFC trial, the participant chose the plural form consistent with the training data ([-ʃ] when the stems ends in [+cons,−voi][+vowel,−back] and [-f] when it ends in [+cons,+voi][+vowel,+back]). The fixed-effects variables are whether the participant is in the experimental or control group (Group), whether the test stem ends in [+cons,+voi][+vowel,+back] or [+cons,−voi][+vowel,−back] (Type), and the interaction between these two variables (Group × Type). The random effects component contained random by-participant and by-item intercepts.

The reference for Group was *control* and the reference for Type was [+cons,−voi][+vowel,−back]. We centered the predictor variables using sum coding to ensure the contrasts sum to zero, and the intercept reflects the grand mean.

The estimates, standard errors, and p-values for the fixed-effects component of the model are shown in Tab. 6.5. There is a significant effect for the interaction between Group and Type. Moreover, an ANOVA test comparing the model to its subset without the interaction shows that the interaction between Group and Type contributes significantly to the model fit ($\chi^2(1) = 15.10, p = 0.00010$).

Having established a significant effect for the interaction between Group and Type, we performed a pairwise Z-test for the Estimated Marginal Means (EMM) between [+cons,−voi][+vowel,−back] and [+cons,+voi][+vowel,+back] for each Group using the *contrast* function of the *emmeans* package in R. We found (Tab. 6.6) a significantly higher rate of choosing the training-consistent form for [+cons,−voi][+vowel,−back] items compared to [+cons,+voi][+vowel,+back] items for both the control group ($p < 0.0001$) and experimental groups ($p = 0.0539$). Since

Table 6.6: Pairwise Z-tests for the Estimated Marginal Means between [+cons,−voi][+vowel,−back] and [+cons,+voi][+vowel,+back] for each Group.

Group	Difference	Est.	Std. Err	p-value
Control	[+cons,−voi][+vowel,−back] − [+cons,+voi][+vowel,+back]	1.955	0.232	< 0.0001
Exp	[+cons,−voi][+vowel,−back] − [+cons,+voi][+vowel,+back]	0.538	0.279	0.0539

Table 6.7: Pairwise Z-tests for the Estimated Marginal Means between Control and Experimental groups for each Type of train-like test item.

Type	Difference	Estimate	Std. Err	p-value
[+cons,+voi][+vowel,+back]	Control − Exp	−2.032	0.289	< 0.0001
[+cons,−voi][+vowel,−back]	Control − Exp	−0.615	0.305	0.0437

[+cons,−voi][+vowel,−back] items take [-ʃ] in the training data, this indicates a higher rate of overextending [-ʃ] than [-f] for both groups. In turn, this suggests that both groups treat [-ʃ] as the default, but that the size of this default preference is lower for the experimental group.

We also performed a pairwise Z-test for the EMMs between groups for each Type, and found (Tab. 6.7) that the experimental group chose the training-consistent form significantly more than the control group for both [+cons,+voi][+vowel,+back] ($p < 0.0001$) and [+cons,−voi][+vowel,−back] ($p = 0.0437$). The larger p-value for [+cons,−voi][+vowel,−back] items is likely because of the control group's preference for [-ʃ], which is coincidentally the training-consistent choice for such forms. These results suggest that training is effective, as experimental group participants chose the training-consistent form for training-like test items significantly more than the control group did.

6.3.2 Hypothesis Evaluation

Since we established in the prior section that [-ʃ] is treated as the default affix by participants, including those in the experimental group, it is appropriate to analyze our hypotheses in Scenario 1 (§ 6.2.5.1), where the predicted generalizations are repeated in (§ 144).

(144) a. LOCAL
/ʃ/ → [f] / [+vowel,+back] __
Elsewhere [ʃ]

b. CONSERVATIVE
/ʃ/ → [f] / [+cons,+voi][+vowel,+back] __
Elsewhere [ʃ]

c. TIER
/ʃ/ → [f] / [+cons,+voi][*] __
Elsewhere [ʃ]

Since the control group also systematically prefers the [-ʃ] suffix, the various hypotheses make the predictions in (145) regarding the responses.

(145) a. LOCAL predicts that experimental group participants will select the [-f] form for test items ending in [+cons,−voi][+vowel,+back] significantly more than the control group does, and will not select the [-ʃ] form for test items ending in [+cons,+voi][+vowel,−back] significantly less than the control group does.

b. CONSERVATIVE predicts that experimental group participants will not select the [-f] form for test items ending in [+cons,−voi][+vowel,+back] significantly less than the control group does and will not select the [-ʃ] form for test items ending in [+cons,+voi][+vowel,−back] significantly less than the control group does.

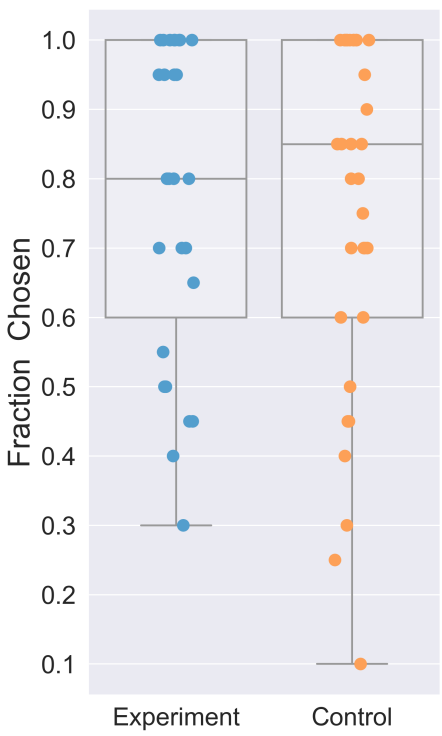
c. TIER predicts that experimental group participants will select the [-f] form for test items ending in [+cons,−voi][+vowel,+back] significantly more than the control group does, and will select the [-ʃ] form for test items ending in [+cons,+voi][+vowel,−back] significantly more than the control group does.

d. STATROBUST predicts that none of the previous three hypotheses predictions will be borne out.

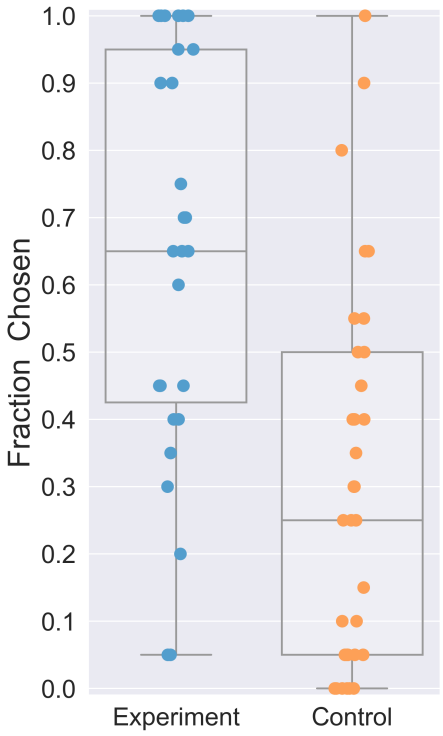
The distribution of [-ʃ] and [-f] responses to each type of novel test item is shown in Fig. 6.4, where both groups choose the [-ʃ] form at high rates for [+cons,+voi][+vowel,−back] forms (Fig. 6.4a), but the experiment group chooses the [-f] much more than the control group for [+cons,−voi][+vowel,+back] forms (Fig. 6.4b). Visually, these results are consistent with the LOCAL predictions.

To test these predictions statistically, we fit a second mixed-effects logistic regression model to the responses to novel test items—i.e., those that are pertinent to establishing which hypothesis the results support. The model was similar to that in § 6.3.1. The categorical, response variable was whether the participant chose the non-default [-f] form on a particular 2AFC trial. The fixed-effects component again contained the variables Group and Type, as well as their interaction. However, the variable Type had the levels [+cons,+voi][+vowel,−back] and [+cons,−voi][+vowel,+back] instead of [+cons,−voi][+vowel,−back] and [+cons,+voi][+vowel,+back]. The reference level was [+cons,+voi][+vowel,−back]. The random effects component still contained random by-participant and by-item intercepts.

The fixed-effects component is shown in Tab. 6.8. The significant interaction between Group and Type indicates that the two groups respond differently to test items. An ANOVA test comparing



(a) Fraction of responses to [+cons,+voi][+vowel,−back] test items where [-ʃ] form chosen



(b) Fraction of responses to [+cons,−voi][+vowel,+back] test items where [-f] form chosen

Figure 6.4: The distributions of which affix form was chosen for novel test items.

Table 6.8: Fixed-effects component of the mixed model fit to responses to novel test items.

Coefficient	Estimate	Std. Err	Pr(> |z|)
Intercept	−1.49597	0.26077	< 0.00001***
Exp	1.12214	0.37761	0.00296**
[+cons,−voi][+vowel,+back]	0.22687	0.07582	0.00277**
Exp × [+cons,−voi][+vowel,+back]	0.80069	0.10386	< 0.00001***

Table 6.9: Pairwise Z-tests for the Estimated Marginal Means between Control and Experimental groups for [+cons,+voi][+vowel,+back] and [+cons,−voi][+vowel,−back].

Group	Difference	Estimate	Std. Err	p-value
[+cons,−voi][+vowel,+back]	Control − Exp	−1.923	0.389	< 0.0001
[+cons,+voi][+vowel,−back]	Control − Exp	−0.321	0.394	0.4151

the model to its subset without the interaction between Group and Type shows that the interaction between these variables contributes significantly to the model fit ($\chi^2(1) = 59.322, p < 0.00001$).

Having established a significant effect for the interaction between Group and Type, we performed a pairwise Z-test for the Estimated Marginal Means between groups for each Type using the *contrast* function of the *emmeans* package. The results for the novel test types are shown in Tab. 6.9. For [+cons,−voi][+vowel,+back] items, the experimental group chose the [-f] form significantly more than the control group, consistent with the LOCAL hypothesis and contradicting the alternative hypotheses. For [+cons,+voi][+vowel,−back] items, the control group did not choose the [-ʃ] significantly more than the control group, which is consistent with the LOCAL hypothesis (as well as the CONSERVATIVE hypothesis, which was ruled out by the last result).

Since the experimental group participants selected the [-f] choice for test items ending in [+cons,−voi][+vowel,+back] significantly more than the control group did, and did not select the [-ʃ] choice for test items ending in [+cons,+voi][+vowel,−back] significantly less than the control group did, these results support the LOCAL hypothesis' predictions. We will consider individual participant's response patterns in the next section, when we compare human behavior to that of different computational models.

6.4 Model Comparison

Having established that the results strongly support the LOCAL hypothesis, we next evaluate how well PLP matches human behavior, in comparison to alternative learning models.

6.4.1 Setup

In addition to PLP, we compare to a trigram model (an n-gram model with $n = 3$), using Laplace smoothing to assign non-zero probability to words with unseen trigrams. The choice of $n = 3$ allows the dependency between the suffix and both the stem-final vowel and the stem-penultimate consonant to fall within the model's view. Thus, the trigram model demonstrates what behavior would result from statistically tracking both dependencies.

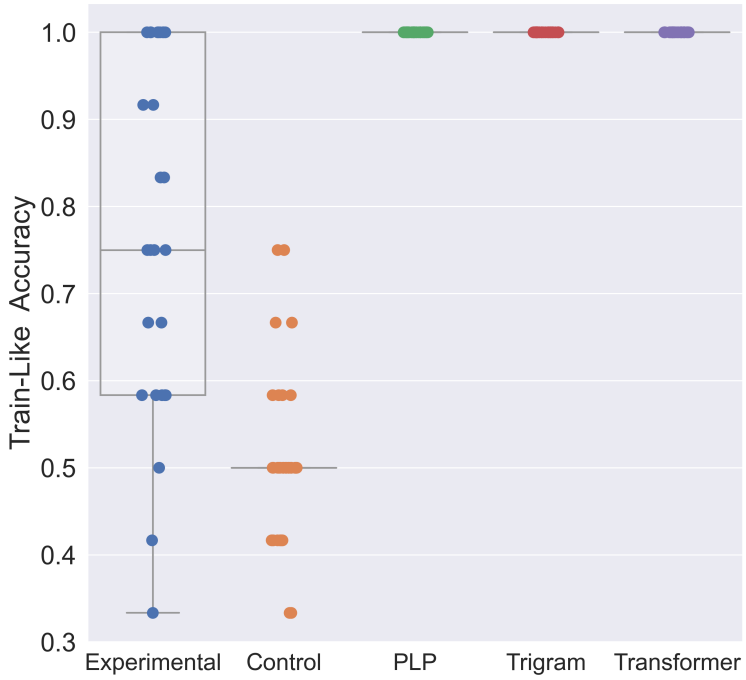


Figure 6.5: Model accuracies on Train-Like Test Items, compared to humans.

We also trained a transformer (Vaswani et al., 2017) language model, following the architecture of BERT (Devlin et al., 2019). We tuned hyperparameters (learning rate $\in [0.0001, 0.01]$, number of epochs $\in \{5, 6, 7, ..., 14, 15\}$, number of attention heads $\in \{1, 2, 3, 4\}$, and number of hidden layers in $\{1, 2, 3, 4\}$) by choosing the combination with the smallest loss on the training-like test items, which the other models do not have access to. This allows the transformer to have access to the entire training dataset for parameter learning, while still allowing us to investigate how the model generalizes to the novel test items, which is the primary question of interest.

We train the models on the same training data as the participants were exposed to. Since human learners appeared to treat [-ʃ] as the default, and since PLP learns to map underlying to surface forms, we treated /-ʃ/ as the underlying affix. The trigram and transformer models are phonotactic models, so the underlying affix is irrelevant for them. We ran each model 30 times with different random seeds. However, given the same training data, PLP and the trigram model are deterministic, so the results for those models are equivalent to those from running a single simulation.

6.4.2 Results

The models' accuracies are shown in Fig. 6.5. All three models achieve perfect accuracy generalizing from the training data to the train-like test items.

Since all models achieve perfect train-like performance, we chose to compare the models to only

the human learners who exhibited strong evidence of learning a training-consistent generalization. This allows us to isolate the question of whether the models generalize beyond the training data to the novel test items in the same way as humans, while ignoring human learners who failed to learn a generalization (e.g., due to not understanding the task or losing focus during it). We follow McMullin and Hansson (2019) in describing successful learners as those who achieve high enough accuracy generalizing to training-like test items to be confident that such performance is not due to chance. Specifically, we selected learners who chose the training-consistent form for training-like test items at a rate greater than the 99% confidence level of a one-tailed binomial test. Because there are only 12 train-like items, this threshold amounts to choosing the training-consistent item at least 11/12 times ($p = 0.9968$).

The models' and successful human learners' rates at choosing the [-ʃ] form for [+cons,+voi][+vowel,−back] items is shown in Fig. 6.6a, and their rate at choosing the [-f] form for [+cons,−voi][+vowel,+back] items is shown in Fig. 6.6b.

The output of PLP is (146).

(146) /ʃ/ –> [f] / [+back] __

As a result, it predicts [-ʃ] for [+cons,+voi][+vowel,−back] words, and [-f] for [+cons,−voi][+vowel,+back] words, which end in a [+back] vowel. Like PLP, successful human learners nearly always selected [-ʃ] for [+cons,+voi][+vowel,−back] items and [-f] for [+cons,−voi][+vowel,+back] items. There is one human outlier in Fig. 6.6b, who we discuss below. The symmetry between humans and PLP is of course expected, since the experiment was designed specifically to test the generalization predicted by PLP, and the hypothesis predictions were supported by the experimental results (§ 6.3).

The one human outlier in Fig. 6.6b consistently picked [-ʃ] for [+cons,+voi][+vowel,−back] items (Fig. 6.6a), but then also frequently picks [-ʃ] for [+cons,−voi][+vowel,+back] items, despite achieving high accuracy on the training-like items. However, this participant did only choose the training-consistent form 11/12 times. Their one mistake was overextending [-ʃ], and they chose the [-ʃ] form for 88% of all test items. It may be that this participant did not learn a generalization, treated [-ʃ] as the default, but got lucky in their performance on train-like test items. Alternatively, this learner may have followed a different generalization strategy than all others, such as CONSERVATIVE.

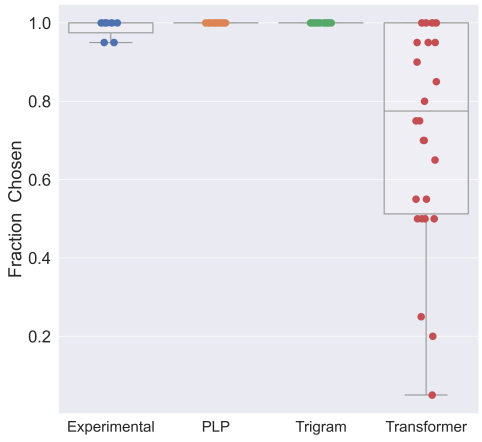
The trigram model was able to achieve perfect train-like test accuracy because, for these test items, the trigrams for the training-consistent choice always have higher probability than the training-inconsistent choice. However, neither of the word-final trigrams in the novel test items occur in the training data (by design, as a poverty-of-stimulus paradigm). These trigrams are thus assigned an equal amount of non-zero probability via Laplace smoothing. When both 2AFC test items received the same score from the tirgram model (which is always the case for novel test items),

we treated the [-ʃ] form as the choice, since participants treated it as a default. This is, by chance, the same choice as experimental participants for [+cons,+voi][+vowel,−back] items (Fig. 6.6a), but the wrong choice for [+cons,−voi][+vowel,+back] items (Fig. 6.6b).

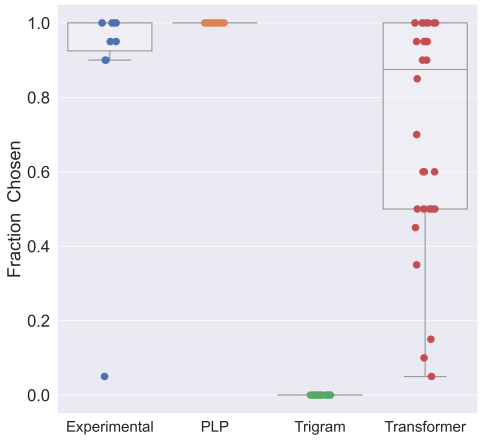
The transformer model behaves very strangely. The fraction of [-ʃ] choices for [+cons,+voi][+vowel,−back] items and [-f] choices for [+cons,−voi][+vowel,+back] items varies widely across the entire range [0, 1] for both, though a majority of model simulations seem to cluster around 1.0 or 0.5. We investigated the choices of each simulation, and found two dominant generalization strategies by the model. On some simulations, the model seems to generalize as humans do, choosing [-ʃ] consistently for [+cons,+voi][+vowel,−back] items and [-f] consistently for [+cons,−voi][+vowel,+back] items. For 16 of the 30 simulations, this correctly predicted the model's choice at a rate greater than the 99% confidence interval of a one-tailed binomial test. However, for another large set of simulations, the model seems to consistently generalize based on the fourth segment in the stem. Thus, for CV.CV items, the model chooses [-f] when the final vowel (fourth segment) is [+back] and [-ʃ] when it is [−back], while for CVC.CV items, the model chooses [-f] when the penultimate consonant is [+voi] and [-ʃ] when it is [−voi]. Using the same 99% confidence interval criteria, this appears to be the generalization strategy of 10 of the 30 simulations. Of the remaining 4 simulations, two appears to generalize based on the penultimate consonant, one consistently selects [-ʃ], and one consistently selects [-f] for CV.CV items, but generalized based on the penultimate consonant for CVC.CV items.

Humans showed no evidence of generalizing in most of the peculiar ways that the transformer language model does. Deep neural networks are known to not generalize well outside of the training distribution. In this case, the novel test items do not follow the training distribution, by design. It is precisely this fact that allows the test items to reveal information about how learners generalize. The fact that many simulations of the transformer model generalize in seemingly peculiar ways (e.g., using the fourth segment of the stem) may make some sense through the lens of the spline theory of deep neural networks proposed by Balestriero and Baraniuk (2018, 2020). In this theory, "a [deep network] constructs a set of signal-dependent, class-specific templates against which the signal is compared via a simple inner product." It may be that different initializations of the network lead to different templates for the [-ʃ] and [-f] classes, and the generalization behavior of each simulation is the result of these templates.

To quantify the quality of fit between the models' generalization behaviors and that of humans, we computed the average rate of choosing [-f] for [+cons,−voi][+vowel,+back] items and that of choosing [-ʃ] for [+cons,+voi][+vowel,−back] items, for each model and for humans. We report the absolute difference between each model's average and humans' average in Tab. 6.10. PLP provides the closest fit.



(a) Fraction of responses to [+cons,+voi][+vowel,−back] test items where [-ʃ] form chosen



(b) Fraction of responses to [+cons,−voi][+vowel,+back] test items where [-f] form chosen

Figure 6.6: The distributions of which affix form was chosen for novel test items.

Table 6.10: The difference between humans' choices and models' choices. The first two columns show the absolute difference in average rate of choosing the [-f] form for [+cons,−voi][+vowel,+back] items and the absolute difference in average rate of choosing the [-ʃ] form for [+cons,+voi][+vowel,−back] items. The third column summarizes these for each model by averaging the first two columns.

Model	[+cons,−voi][+vowel,+back]	[+cons,+voi][+vowel,−back]	Average
PLP	**0.1136**	**0.0136**	**0.0636**
Trigram	0.8864	**0.0136**	0.4500
Transformer	0.1680	0.2514	0.2097

6.5 Disucssion

The results in § 6.3.1 demonstrated that participants in both the control and experimental groups treated [-ʃ] as the default affix, and that participants in the experimental group constructed a generalization from the training data that allowed them to accurately predict the plural form of train-like test singulars. This conforms to the interpretation that the experimental-group participants constructed one of the generalization described in rule-notation in (139) and repeated in (147), which treat [-ʃ] as the default.

(147) a. LOCAL
/ʃ/ → [f] / [+vowel,+back] __
Elsewhere [ʃ]

b. CONSERVATIVE
/ʃ/ → [f] / [+cons,+voi][+vowel,+back] __
Elsewhere [ʃ]

c. TIER
/ʃ/ → [f] / [+cons,+voi][*] __
Elsewhere [ʃ]

The results in § 6.3.2 strongly support the LOCAL generalization strategy, as participants in the experimental group systematically chose the [-f] form for novel test singulars that ended in back vowels but voiceless consonants (i.e., [+cons,−voi][+vowel,+back]-final stems). The LOCAL generalization (147a) predicts precisely this behavior, while the CONSERVATIVE and TIER generalizations (147b) predict that learners would choose the default [-ʃ] for such items, which fall under the Elsewhere conditions of (147b)-(147c).

Since the alternation expressed in this artificial language was arbitrary and lacked phonetic motivation, these results also corroborate other experiments that have demonstrated humans can learn arbitrary phonological rules (Seidl and Buckley, 2005; Beguš, 2018).

It was neither obvious nor surprising that participants would treat [-ʃ] as the default affix. We chose this suffix because of the saliency of fricatives in word-final position (e.g., compared to stops or nasals) and the fact that the allomorphs had to agree in voicing in order to ensure that the voicing of the penultimate consonant did not interact with the alternating affix. Since neither [-f] nor [-ʃ] is an English affix, there was a chance that participants did not treat either as the default. However, the results strongly suggested that [-ʃ] was treated as the default. This is not surprising, given the higher saliency of [ʃ] compared to [f] (Behrens and Blumstein, 1988) and the similarity between [-ʃ] and the English plural [-s]. Future work could confirm that the LOCAL hypothesis holds up when one suffix is not treated as the default. This could potentially be accomplished by using liquids [-l]

and [-r], which are still clearly differentiated but do not bear a resemblance to the English plural affix.

6.5.1 Algorithms and Formal Language Theory

Formal-language-theoretic approaches to phonology are often interested in characterizing the amount of computational expressivity needed to cover all typologically-attested phonological generalizations, while simultaneously being maximally-restrictive (Heinz, 2018, sec. 4). One primary motivation for doing so is to restrict the hypothesis space of a phonological learner, thereby making learning easier. At present, many working in this domain believe that phonology is likely *subregular*, meaning that the most restrictive class of the Chomsky (1956) hierarchy, which is sufficiently expressive (Johnson, 1972; Kaplan and Kay, 1994), is also *overly* expressive. This hypothesis is called the *subregular hypothesis*.

A consensus has not been reached on how restrictive phonological theories can be before being too restrictive, but some classes of generalizations appear to certainly be necessary. One is the class of k-input-strictly-local (kISL) (Chandlee, 2014) string-to-string functions, which capture input-output mappings where the output for each input segment can be determined by referencing a fixed-size window of length k around the input segment. 2ISL processes characterize the wide range of processes involving adjacent dependencies, and 3ISL processes include phenomena like intervocalic voicing and English flapping (Chandlee, 2014, sec. 6.3.1.2). The rules constructed by PLP can be interpreted as ISL (§ B.1). A second class is the class of k-output-tier-strictly-local (kOTSL) string-to-string functions, which cover a large number of long-distance phonological processes (Burness and McMullin, 2019; Burness et al., 2021); the rules constructed by D2L can be interpreted as OTSL (§ A.1.2).

The rules in (139), which describe possible generalizations consistent with the training data, are repeated once more in (148).

(148) a. LOCAL
/ʃ/ → [f] / [+vowel,+back] __
Elsewhere [ʃ]

b. CONSERVATIVE
/ʃ/ → [f] / [+cons,+voi][+vowel,+back] __
Elsewhere [ʃ]

c. TIER
/ʃ/ → [f] / [+cons,+voi][*] __
Elsewhere [ʃ]

The first two rules (148a)-(148b) are 2ISL and 3ISL, respectively. In the first, the surface realization of /-ʃ/ is determined by a sequence of fixed length 2, containing /-ʃ/ and the final vowel of the stem; in the second, it is determined by a sequence of length 3, which also includes the stem-penultimate consonant. Since [ʃ] and [f] are consonants, the third rule (148c) is 2ISL on a [+cons] tier. In other words, it is tier-strictly-local (2TSL).

In summary, since all three generalizations fall within classes of string-to-string functions that appear to be lower-bounds for the expressivity of phonology, their formal-language-theoretic characterizations predict them all as possible rules. In contrast, the algorithmic characterization of PLP and D2L allows them to make a precise prediction about what generalization will be constructed in this learning scenario. This is an example of how our algorithmic approach to phonology provides predictions regarding human generalization that complement formal-language-theoretic characterizations of phonological generalizations.

6.5.2 Future Directions

The artificial language alternation constructed for this study intentionally lacked phonetic motivation because we wished to control for the possibility that adjacent interactions (e.g., voicing assimilation) may be more phonetically-motivated than non-adjacent interactions. A more challenging test of the LOCAL hypothesis would be a condition where the adjacent dependency lacks phonetic motivation, while the non-adjacent dependency has clear phonetic motivation. For example, an artificial language could be constructed in which bounded sibilant harmony leads a sibilant suffix to alternate between [-s] and [-ʃ] to match the anteriority of the final sibilant in SV-final stems. This non-adjacent dependency has the phonetic motivation common to sibilant harmony. If the backness of the stem-final vowel also fully predicts the sibilant (e.g., back-vowel-final stems take [-s] and front-vowel-final stems take [-ʃ]), then the adjacent dependency lacks the clear phonetic motivation that the non-adjacent dependency enjoys. If learners still generalize consistently with the LOCAL hypothesis, this constitutes even stronger evidence in favor of the hypothesis.

CHAPTER 7

Conclusion and Discussion

7.1 Main Results and Contributions

In this dissertation, we have built an algorithmic, learning-based approach to phonology. In comparison to prior work on phonological learning, we have placed emphasis on the learning *algorithm*, which we have built bottom up by identifying already-established psychological mechanisms. The emphasis on the process of learning instead of a supposed end-state has led to a number of results and contributions to the study of phonology. We believe these insights improve our understanding of the processes by which humans construct phonological systems, and suggest how rich and effective generalizations can be built from limited input. These results, which consistently out-perform neural network models on small amounts of data, suggest a line of research, beyond ever-larger language models, toward high-quality language technology for speakers of all the world's languages. We now review the main results and contributions.

7.1.1 Adjacency First

In § 2.4, we identified humans as having a sensitivity to adjacent dependencies, and that sensitivity to non-adjacent dependencies emerges later in development and is only resorted to when adjacent dependencies become comparably weak. This became the core building block of our computational models D2L (§ 3) and PLP (§ 5). In § 5, we showed that PLP accounts for the fact that more local phonological processes are easier for humans to learn than less local processes (Baer-Henney and van de Vijver, 2012). More strongly though, our models predicted that learner's do not track all dependencies and then attend to the relevant ones. Rather, they predicted that if adjacent dependencies allow for accurate prediction, then learners will effectively be blind to any dependencies beyond adjacency. We confirmed these predictions in our human-subject experiment in Chapter § 6, which contradicts the predictions made by most statistical and neural models.

These chapters also highlight the importance of the learning *algorithm*. Traditional approaches to phonological learning place the structures learned at the starting point of inquiry, and only move

towards how these structures are learned once a theory of the structures is developed. Such theories can characterize the difference between local and non-local generalizations, but do not give an explanation for why one is constructed over the other when each is equally descriptively adequate.

7.1.2 Change in Representation as the Consequence of Adjacency First

In Chapter § 3, we demonstrated how non-flat representations of words (tiers) are a natural consequence of a learning model grounded in humans proclivity for tracking adjacent dependencies. Our model D2L, learns phonological processes directly over the flat, input representation when no change of representation is needed, but changes representation when needed. This iterative change of representation when and only when needed unifies local and non-local learning because local alternations become a special case where no change of representation is needed. This become especially clear in § 4, when D2L accounted for both non-local vowel harmony and local voicing assimilation in Turkish.

A second implication of the iterative change of representation is that D2L accounts for cross-linguistic variation in *what* change of representation is needed, by demonstrating that blockers are included on the changed representation and neutral segments are removed.

7.1.3 Productivity and Lack Thereof

Combining the models from § 4 and § 3 provided a process by which a learner can go from morphologically-analyzed surface forms to productive knowledge of Turkish vowel harmony by the time that the learner's vocabulary reaches 1K words. This accounted for the behavior of children, who extend vowel harmony productively to loan and nonce words at a young age (Altan, 2009), when their vocabularies likely contain less than a thousand words.

However, the same § 4 model gave a possible learning-based explanation for the apparent *lack* of productive knowledge of voicing alternations in Dutch noun paradigms. Our model suggests that children may lack this productive knowledge because they are able to form effective morphological generalizations without it. This provides a novel interpretation of this instance of a lack of productivity, and indicates the potential pitfalls of inferring general principles about the nature of phonological knowledge—e.g. that it is usage-based (Kerkhoff, 2007)—from results pertaining to a single morphophonological process.

7.1.4 The Typologically Rare

Chapter § 5 allows for a simple account of how processes that lack or operate in opposition to phonetic motivation can be learned. This was shown of the well-studied case of post-nasal voicing in

Tswana, which has been shown to be productive (Coetzee and Pretorius, 2010) despite operating against apparent phonetic motivation (Beguš, 2018). This result is consistent with prior experimental results showing that humans readily learn such processes (Seidl and Buckley 2005, Beguš 2018, ch. 6), even if they are typologically rare.

This result is further corroborated by our own experiment in Chapter § 6, where participants readily learned phonological generalizations, despite the artificial language's morphophonological process lacking any apparent phonetic motivation. We discuss our view on the relationship between typology and our approach to phonology in § 7.3.4.

7.2 A Critical Comparison with Neural Networks

The artificial neural network (henceforth neural network) is usually traced to the neuron model of McCulloch and Pitts (1943), who studied how networks of logic gates could compute logical functions. This work influenced Von Neumann (1945)—indeed, modern computers still contain networks of logic gates—and it was McCulloch and Pitts who suggested that something like Turing (1937)'s formulation of computation could provide a computational theory of mind (Rescorla, 2020). Nevertheless, neural networks were taken up as the primary computational models of *connectionism* (McClelland et al., 1986, 1987), which is often viewed as a rival of Turing-style computation within the computational theory of mind (Rescorla, 2020). Beyond their role in cognitive science, neural network models have seen increasing success in natural language processing tasks.

We will briefly review (§ 7.2.1) some of the arguments against neural networks as models of the mind,[1] and will then (§ 7.2.2) argue that the training procedure used for neural networks obscures the learning *process* in comparison to our algorithmic approach, which allows greater insight into the processes by which the mind might construct a phonological system. We will then turn to the issue of training data size (§ 7.2.3) and generalization beyond the training distribution (§ 7.2.4). We will argue that our results, in comparison to those of neural networks, are more consistent with cross-linguistic evidence relating vocabulary size to linguistic development, and are better aligned with human behavior on test data outside the training distribution. We will then briefly discuss aspect of interpretability (§ 7.2.5). Lastly, we will provide indicators of how our approach's emphasis on learning *algorithms* and success on small training sizes could contribute to the practical use of neural networks in speech-related technologies (§ 7.2.6).

[1] See Rescorla (2020) for a more detailed discussion.

7.2.1 Neural Networks as Computational Theories of Mind

The classical computational theory of mind approached the study of mind by using the Turing (1937) model of computation as a theory of how the mind functions. The Turing model of computation makes an explicit distinction between *data* (memory)—stored on an "external" tape—and *algorithm*—the machine's finite table of instructions. Neural network models do not in general make such a distinction: information is carried forward in time through the network parameters, such that data and algorithm are blurred into a single network structure (Gallistel and King, 2011).

The neural network has become the primary model for the *connectionist* computational theory of mind. Some connectionists take neural networks to provide a fundamentally distinct model of the mind from that of the Turing model. This position is sometimes called *eliminative* connectionism (Pinker and Prince, 1988; Rescorla, 2020) because, under this view, if the mind is appropriately modeled as a neural network, then it eliminates the relevance of the Turing model in the study of mind. An alternative position is *implementationist* connectionism (Pinker and Prince, 1988; Rescorla, 2020), which views neural networks as implementing classic Turing-style computation. Under this view, both neural network computation and Turing computation are appropriate levels of modeling.

Several arguments have been leveled against connectionism, some leaving the door open to implementationist connectionism, and others rejecting connectionism outright. Gallistel and King (2011) argue that Turing computation is a better model of the mind than are neural networks. Much of their argument comes down to the distinction mentioned above: the Turing machine has an explicit structure for memory, separate from the program, while neural networks do not. Gallistel and King argue that even apparently simple forms of cognition, like insects navigating via dead reckoning,[2] requires operating over memory in a way that is trivial in a Turing model of computation but implausible in a neural network model.

Furthermore, reconsidered research from the early 20th century (Gershman et al., 2021) together with research from the last few years (Johansson et al., 2014; Bédécarrats et al., 2018) strongly suggests that individual cells are capable of learning and retaining learned information, which implies that learning and memory cannot be solely attributed to *networks* of interconnected neurons.

In a famous critique of connectionism, Fodor and Pylyshyn (1988) argued that human language exhibits *systematicity*, which is a property in which the actuality of certain mental states entails the possibility of other mental states. The classic example is that if someone understands the sentence "John loves Mary," then this implies that they must also be able to understand the sentence "Mary loves John."[3] Fodor and Pylyshyn demonstrate that systematicity is not a property of neural networks unless they implement the classic Turing style of computation.

[2] Also called *path integration*.

[3] Regardless of their attitudes towards these propositions.

While we find these arguments compelling and feel that they have yet to be satisfactorily addressed by connectionism, the use of neural networks as computational models of the mind persists (e.g., Joanisse and McClelland 2015). Regardless of the conclusion about whether neural networks are plausible as computational theories of the mind, we believe that our algorithmic approach to phonological learning is in many cases better-suited to understanding the processes by which the mind constructs phonological systems. We turn to this next.

7.2.2 Learning Algorithms and Developmental Predictions

We would like to suggest a distinction between a neural network as a computing *device* and the *procedure* used to train it. Some neural network architectures have been shown to be Turing Complete under certain conditions, including recurrent neural networks (Siegelmann and Sontag, 1992), Neural Turing Machines (Graves et al., 2014), and Neural GPUs (Kaiser and Sutskever, 2016). Moreover, properties like *systematicity*, discussed in the section above, are possible in neural networks if they implement Turing-style computation. However, these results pertain to the functions expressible with neural networks, and not the process by which such functions might be learned. Indeed, neural networks are usually trained via *back propagation*, which is not usually taken to be a hypothesis about the actual process of learning because it appears to be impossible in actual neural networks (Crick et al., 1986; Rescorla, 2020). Thus, even if neural networks are taken to be good models of linguistic knowledge—which is itself a controversial idea (see above)—they do not provide a hypothesis about the *process* by which that knowledge may be constructed. In short, neural networks are not learning algorithms.

To make this point concrete, consider the issue of developmental predictions. In a study that has featured prominently in cognitive science, Rumelhart and McClelland (1986) used a recurrent neural network, which, as we discussed above, are Turing Complete under some conditions, to model learning of the English past tense. The acquisition of the English past tense is well-known to follow a *U-shaped* trajectory, in which children initially produce the form of irregular verbs correctly (e.g., *run* ~ *ran*), but fall into a period of overregularization (e.g., *run* ~ *runned*) before eventually returning to correct production of irregulars. The period of overregularization thus leads to a drop in the frequency of correct past-tense forms for irregular verbs, between two peaks; hence the U-shape. This is interpreted as evidence of rule-like behavior, since the onset of overregularization marks the point of acquisition of a productive, regular past tense formation rule that children initially overextend to irregulars. Rumelhart and McClelland claimed that this characteristic U-shape curve emerges from the training of their neural network, but achieved this result by manipulating the order of training data so as to all but guarantee the result. If one runs the same experiment with a modern RNN and without so manipulating the order, as Kirov and Cotterell (2018) did, no U-

shape curve emerges. It was thus the order in which the data was presented—effectively a learning algorithm separate from the RNN—that led to the developmental prediction, not the RNN itself.

In contrast, our models make clear developmental predictions. For instance, our models clearly predict that learners will track adjacent dependencies *before* changing representation or expanding attention. This temporal ordering of events is critical to making predictions about development and, by that very fact, allowing for intelligibility into the developmental process. It is this iterative component of our learning algorithms that predicted that the participants in § 6 would show no sensitivity to non-adjacent dependencies given that adjacent dependencies were equally statistically robust.

Moreover, in § 5.4 we showed that PLP matched Berko (1958)'s developmental finding that children show productive knowledge of English voicing alternations before they do for alternations involving epenthesis. And in § 4, we showed how our model for learning underlying forms predicts some processes, like vowel harmony in Turkish, to be productive and others, like voicing alternations in Dutch, not to be. These developmental predictions are substantiated by experimental studies in Turkish (Altan, 2009) and Dutch (Zamuner et al., 2006, 2012).

The conceptual separation between how knowledge is represented (e.g. rules or neural networks) and the process by which it is constructed also opens the possibility of applying the learning from our work to improve neural network approaches to phonology, which we turn to in § 7.2.6.

7.2.3 Accuracy on Small Data

Cross-linguistic acquisition studies show that children have acquired much of their morphophonology by the time they are 3 years old (Lignos and Yang, 2016; Kodner, 2022). This is true for English (Brown, 1973), as well as languages with a greater degree of morphological inflection (Aksu-Koç and Slobin, 1985; Deen, 2005; Caprin and Guasti, 2009). At this young age, children have very small vocabularies. The size of children's vocabularies varies a fair amount child-to-child, but consistently falls within the range of a few hundred to a thousand words, regardless of the child's native language (Anglin, 1993; Fenson et al., 1994; Hart and Risley, 1995; Bornstein et al., 2004; Szagun et al., 2006).

Thus, any computational theory of child language acquisition must at a minimum achieve comparable results on a training set of comparable size. In every natural-language experiment in this dissertation, our proposed model achieved higher accuracy at every training size than the neural network comparison model(s). We summarize these results in Tab. 7.1. On training sizes of no greater than 1K words, our proposed models achieves over 0.9 accuracy on every natural language experiment in the dissertation, while neural network comparison models never do.

Recently, some have taken interest in the potential of transformer-based language models trained

Table 7.1: Summary of neural network results in comparison to our proposed models.

Dataset	Training Size	Our Model's Acc	NN Acc
Latin (Ch. § 3)	97	0.965 ± 0.03	0.774 ± 0.11
German (Ch. § 5)	400	1.000 ± 0.00	0.543 ± 0.04
Dutch (Ch. § 4)	700	0.940 ± 0.02	0.878 ± 0.26
Turkish (Ch. § 3)	958	0.984 ± 0.01	0.725 ± 0.18
Finnish (Ch. § 3)	975	0.980 ± 0.01	0.816 ± 0.03
Turkish (MorphoChallenge) (Ch. § 4)	1000	0.917 ± 0.05	0.647 ± 0.15
Turkish (CHILDES) (Ch. § 4)	1000	0.924 ± 0.09	0.741 ± 0.20

on considerably smaller sizes than the behemoths like GPT (e.g., Hosseini et al. 2022). However, calls like the *BabyLM Challenge* chose 100M word tokens as the target size, because they estimated that this approximates the amount of exposure of a child prior to age 12 (Warstadt et al., 2023). The dataset contains around 1M word types, consistent with Heap's law of type-token ratios discussed in § 2.2. In light of the empirical studies discussed above, this likely over-estimates the amount of input that a child needs to acquire morphophonology by 3-4 orders of magnitude (Kodner, 2022). Warstadt et al. also released a 10M word dataset, which contains over 200K word types, still over-estimating the amount of input by 2-3 orders of magnitude. Thus, any success with models trained on these datasets does not contradict the comparatively low accuracy of neural models in this dissertation.

Beyond the importance of high accuracy on small training data for developmental plausibility, these results show promise for natural language processing (NLP) and automatic speech recognition (ASR) for a vast majority of the world's languages, where large datasets are unavailable. We discuss promising directions in § 7.2.6.

7.2.4 Non-Human-Like Generalization Beyond Training Distribution

In Chapter § 6, we compared the generalization behavior of computational models to that of humans in our artificial language experiment. The experiment was a poverty-of-stimulus paradigm, which means that there are multiple generalizations consistent with the training data, and also that the test data does not come from the training distribution. When we compared a transformer language model, we observed that the model could achieve perfect accuracy on in-training-distribution items, but behaved in strange ways on the out-of-training-distribution items. For instance, on 10 simulations, the model appeared to make its prediction in the 2-alternative forced-choice (2AFC) test based on the fourth segment in the word. No human participant showed evidence of generalizing in this way.

Neural networks have been shown to easily fit randomly-labeled data, which indicates that they are capable of memorizing the training dataset (Zhang et al., 2021) even those as large as ImageNet (Deng et al., 2009). Since neural networks nevertheless generalize from non-random labels, Zhang et al. concluded that their generalization cannot be attributed to traditional statistical concepts like regularization of an over-parameterized model. This result makes it difficult to interpret what it means for a neural network to generalize. We have found the spline theory of deep neural networks proposed by Balestriero and Baraniuk (2018, 2020) to be helpful. We repeat the quote from § 6.4, which clearly states the idea: "a [deep network] constructs a set of signal-dependent, class-specific templates against which the signal is compared via a simple inner product." In our interpretation, the fact that a neural network model was able to achieve perfect accuracy on the training-like items for our experiment while behaving oddly on the novel test items may be because the model was able to construct class-specific templates that the train-like items were mapped to the same template as their related training counterparts, while novel items were mapped in peculiar ways. Regardless, the model's behavior suggests that neural networks are not systematic in the way that they generalize beyond the training distribution in artificial language experiments. The same can not be said of the human participants.

7.2.5 Interpretability

The direct output of our models are interpretable rules, readable by a linguist with no extra training. Interpretability is an active area of research in neural networks, but the techniques require a great deal of architecture-specific mathematical sophistication beyond what may be directly accessible to most linguists. For example, Hewitt and Manning (2019) propose probes to identify syntactic tree structure implicit in the vector geometry of language models, but the probes require knowledge of linear algebra, which is not part of the standard education of a linguist. In the domain of phonology, Beguš (2022) requires sophisticated manipulation of Generative Adversarial Network latent variables—a process described in Beguš (2020)—to identify rule-like constructs implicit in the model.

7.2.6 Potential Contributions to Sample-Efficient NLP and ASR

There are over 7,000 languages in the world (Eberhard et al., 2023), and many of these have no written form, and of those that do, only a handful have readily-available large datasets. Such languages are called *low-resource languages*. In order to provide language technologies for speakers of most of these languages, it is important to be able to generalize rapidly and accurately from small amounts of data that is as close as possible to spoken form. Automatic speech recognition for *low-resource languages* (i.e. languages for which large datasets are not available) is an active area

of research (Besacier et al., 2014; Reitmaier et al., 2022), as is language modeling over speech for languages and contexts lacking text (Lakhotia et al., 2021; Polyak et al., 2021; Kharitonov et al., 2022).

Our models operate over representations of spoken words as sequences of segments, which are in turn feature sets of distinctive features. However, in future work we will consider how these levels of representation are constructed from the auditory signal (see § 7.3.2). The fact that our models consistently achieve higher accuracy than neural models on datasets containing hundreds to thousands of words (see discussion above in § 7.2.3), indicates that our approach shows great promise for providing an alternative to neural-based approaches to NLP and ASL when large datasets are unavailable.

We will now discuss how the insights of our models could be used to inform the design of neural architectures and training procedures to make them more sample efficient.

7.2.6.1 Iteratively-Expanded Attention

While convolutional architectures are highly efficient in language modeling and seq-to-seq tasks (Gehring et al., 2017), their downside is that they can only capture dependencies within a fixed distance. In contrast, transformer architectures (Vaswani et al., 2017) are capable—at least in principle—of capturing arbitrarily-distant dependencies. In this dissertation, we have demonstrated the value of starting with adjacent dependencies and iteratively expanding outward only when needed (chapter § 5). This suggests a possible extension to convolutional-based language modeling that could strike a balance between efficiency and expressivity.

Like, PLP, a convolutional layer has a parameter, usually called its *kernel size*, that determines the width of its attention. Unlike PLP, this value is fixed prior to training, either by a practitioner's decision or by hyperparameter tuning.

Moreover, if convolutional layers are stacked, information from longer distances is able to cascade through the layers. For instance, if first-layer hidden representation $h_i^{(1)}$ is dependent on input representations $\{x_{i-1}, x_i, x_{i+1}\}$[4] and second-layer hidden representation $h_i^{(2)}$ is dependent on $\{h_{i-1}^{(1)}, h_i^{(1)}, h_{i+1}^{(1)}\}$, then information about x_{i-2} and x_{i+2} indirectly contributes to $h_i^{(2)}$ via $h_{i-1}^{(1)}$ and $h_{i+1}^{(1)}$. Still, the distance that information can propagate is architecturally bounded. In short, convolutional neural networks have an architecturally-specified width of attention, which can be expanded by increasing a single layer's attention width or by stacking additional convolutional layers.

PLP suggests a novel training procedure for convolutional-based language models, in which a single-convolutional-layer model is trained to convergence and evaluated. If the model's quality fails to meet a satisfactory threshold according to a chosen evaluation metric, then an additional

[4]As would be the case with a kernel size of 3.

convolutional layer could be concatenated and the model trained again to convergence. This process would repeat iteratively—possibly freezing the weights of previously trained layers—until the threshold is met. This is, in effect, a "neuralification" of PLP, where the interpretable linguistic rules are replaced with convolutional layers, and highlights the distinction we made in § 7.2.2 between the way in which linguistic knowledge is represented and the algorithm by which it is constructed.

7.2.6.2 Non-Local Dependencies

Because text is naturally represented as *strings*, language modeling has usually been framed in terms of string-based tasks like next-token or masked-token prediction. The expressive power of language models has been expanded in order to handle dependencies between tokens far away in a string. However, this expressiveness comes at the cost of requiring massive training data, computational resources, and power (Bender et al., 2021). A prevalent theme in linguistic theory is that most dependencies that appear to be long-distance may actually be local on some representation, usually some type of *graph*. Graphs have been used extensively in NLP (Nastase et al., 2015), and recently attention has been given to graph neural networks in particular (Wu et al., 2021). A significant advantage of this approach is that graphs can encode rich, explicit structure, such as dependency parses of sentences. However, this explicit graph structure is often not known in advance. This is the case with autosegmental graphs (tiers) because, as we discussed in Chapter § 3, the graph needed to render dependencies local depends on the alternation. Some works, including Liu et al. (2019); Reddy et al. (2019); Wu et al. (2021), have proposed methods for learning a graph structure while jointly learning over the resulting structure. These approaches use node embedding similarity to construct a thresholded adjacency-matrix. However, the critical component of D2L was its iterative creation of a deletion set, the complement of which constitutes the graph (tier). In § 5.5, we discussed why this leads it to match human behavior in the artificial language experiments (§ 3.4) and succeed at learning natural language alternations (§ 3.5) even when they contain complexities like blocking segments (§ 3.5.3).

The key insight of D2L is that non-locality is *systematic*. For example, while Turkish vowels harmonize across arbitrary numbers of intervening consonants, the intervening segments are always consonants. More generally, in D2L's tier-based generalizations, interacting segments cross arbitrary numbers of intervening segments, but the intervening segments can always be characterized as a single set—the *deletion set*. D2L provides a means of constructing a graph that captures these systematic non-localities by linking only segments that are adjacent after removal of the deletion set. The iterative deletion of segments in D2L could be used in tandem with a graph neural network that learns *locally* over the resulting graph representations of words in order to achieve greater sample efficiency when non-local dependencies are present.

7.3 Limitations and Future Directions

7.3.1 Putting Pieces Together

The models proposed in this dissertation are all related. The model from § 4 constructs, in response to surface alternation, abstract representations of morphemes. These introduce discrepancies between underlying forms and their concrete surface realization, which requires learning morphophonological processes to map between the levels of representation. The models D2L and PLP in chapters § 3 and § 5 provide an account of how these mappings are constructed.

Both PLP and D2L start by tracking adjacent dependencies. However, whereas D2L changes representation when adjacent dependencies are insufficient, PLP expands the width of its attention. Because attention is characterized as a *window* in PLP, the rules it constructs are *input-strictly local* (Chandlee, 2014), which means they predict the surface form of an underlying segment by referencing only segments within a fixed length of it. In contrast, D2L's rules involve a change of representation, and then apply locally over the learned representation. These rules are *tier-strictly local*, which means that they predict the surface form of an underlying segment by referencing only segments within a fixed length on the learned representation, but that representation is flexible enough that dependencies arbitrarily far away from the target segment in the input representation can be cast into a fixed distance of it on the new representation. Consequently, PLP can be thought of as a model for learning processes that are strictly-local over a representation, whether that be an input string representation or a tier representation. However, the order in which these two learning procedures interact needs to be investigated in future work. When adjacent dependencies are not sufficient for generalization, should the model first change representation or first expand the window of attention?

Secondly, the problem of morphological learning needs to be more closely connected to that of learning phonology. Recent work (Cotterell et al., 2015; Rasin et al., 2018; Ellis et al., 2022) has recognized the importance of studying these problems together, focusing jointly on the problems of learning underlying forms, morphological rules, and phonological rules (though with limited results on natural language problems). In our view, the approach we have taken, which breaks the overall learning process into smaller individual processes, allows for greater insight into the algorithms by which humans construct a morphophonological system. However, this cannot be done without a recognition of the fact that the individual problems are intricately connected, and a complete story must consider the problems together. Thus, it will be important for future work to begin to bring the components proposed in these chapters together into a unified learning system.

7.3.2 Lower Levels of Representation

Throughout this dissertation, we have treated the representation of words as made up of discrete phonological segments, which were in turn treated as sets of distinctive features. This is common in phonological theory (e.g., Jakobson and Halle 1956; Chomsky and Halle 1968) and there is strong psycholinguistic (Liberman et al., 1957, 1967; Finley and Badecker, 2009) and neurolinguistic (Chang et al., 2010; Mesgarani et al., 2014) evidence for some representation of this sort. However, how these representations are constructed from a continuous speech signal is not fully understood, and neither is the exact nature and content of segmental representations. If we get the representation of words wrong, we risk the possibility that conclusions drawn from models operating over them may no longer hold upon a correction of the representation. Nevertheless, the aforementioned research suggests that the use of discrete phonological segments made up of features is likely on the right track, even if not completely correct.

In future work, we plan to apply the algorithmic approach to phonology that we have argued for in this dissertation to the question of the nature of lower levels of representations, and how they are constructed. We will seek independent psychological mechanisms, such as sensitivity to particular acoustic features (Mesgarani et al., 2014), that can form the basis of a learning-based account of the units of words.

7.3.3 Variation

At this point, our models learn categorical rules. An important and challenging topic in phonology is variation. As Coetzee and Pater (2011) discuss, rule-based theories of phonology can handle variation by marking a rule optional, and then augmenting it with a probability (Labov, 1969; Vaux, 2008). This approach requires, at a minimum, a statement of the rule's structural description, just as it does for a categorical rule. Thus, while constructing a probabilistic rule requires strictly more information than a categorical one (at minimum, the rule's structural description *and* a probability) the statement of the rule's structural description is shared by categorical and probabilistic rules alike. Future work will thus consider how PLP and D2L can be extended to capture probabilistic information.

In the case of long-distance processes in particular, the variability of some processes is affected by the distance between between segments: the probability of harmony decreases with the distance between harmonizing segments, and blocking is sometimes probabilistic (Hayes and Londe, 2006; Hayes et al., 2009; Bennett, 2013). Mayer (2021) argued that distance decay and gradient blocking can be captured without reference to distance by extending tier-based representations to allow probabilistic projection onto the tier. In future work, D2L could be extended to allow for probabilistic projection, though we also believe that such a solution leaves upon the question of *why* such

projection may be probabilistic.

7.3.4 Typology

Because Optimality Theory (Prince and Smolensky, 1993) assumes that languages differ (phonologically) only in their ranking of a set of universal constraints, the theory makes explicit and quantitative typological predictions. As a result, the rise of OT tightened the connections between phonological theory as a theory of the human capacity for language and phonological theory as a predictor of typology.

Our proposed algorithmic approach to phonology does not make as clear typological predictions. We believe that OT remains a powerful tool for typology, but that a requirement that a theory of phonology must directly predict typology is not justified. When asking why languages have the structure that they do, there are multiple causal pathways to consider, beyond the consideration of what grammars are expressable and learnable by the human mind. The typologist's goal in explaining why observed languages have the structures that they do and not others is similar to the evolutionary biologist's goal in explaining why observed organisms have the structures that they do and not others. While some structures are attributable to predictable processes of evolution, others are the result of historical contingency, which Gould (1989, 1999, 2002) characterizes as outcomes of history that are entirely and in principle unpredictable, but are nevertheless intelligible in retrospect. For example, the lack of five-eyed sea creatures roaming today's seas is probably because the five-eyed *Opabinia* was wiped out by an extinction event sometime in the Cambrian, which cut off evolutionary branches that may have later emerged from it. The reason is fully intelligible in hindsight, but not predictable. Similarly, the typological absence of a given property in languages could in some cases be the consequence of necessary lineages destroyed by geopolitical events, which we would not be reasonable to require be captured by linguistic theory, lest we require linguistics be a Theory of Everything.

This argument is consistent with that of Hale and Reiss (2008, sec. 1.2), who argue for a hierarchy of languages—ATTESTED ⊂ ATTESTABLE ⊂ HUMANLY COMPUTABLE ⊂ STATABLE—in which it is the HUMANLY COMPUTABLE languages that they take theories of Universal Grammar to be about. Thus, in Hale and Reiss's view also, linguistic theories need not map directly onto predictions about attested languages.

Nevertheless, while we reject an immediate mapping from linguistic theory to typological prediction, the two topics are not independent either. For example, McMullin (2016) clearly recognizes a distinction between learnable patterns and attested patterns in the first figure of his dissertation (p. 4), which places the attested languages as a subset of the learnable languages. McMullin then proceeds to provide a clear example of how consideration of the properties of the language learner

does make typological predictions. McMullin demonstrates that a learner restricted to searching tier-strictly-local generalizations cannot form a restriction that requires consonants to harmonize across exactly one intervening consonant, since doing so would require simultaneously including and excluding the set of consonants in the language from the tier—a logical contradiction. The hypothesis that learners are indeed restricted to this class of generalizations then makes the (correct) prediction that no such languages should exist, since they would be unlearnable by such a system.

As explicit and interpretable learning algorithms, our proposed models can make similar predictions. We have not formulated these predictions in the dissertation—in part because of our skepticism about the legitimacy of drawing inferences between typology and phonological theory—but could do so in future work.

7.3.5 Rule Interaction

In § 5.2.3.3, we discussed how PLP handles rule ordering. An important area for future research will be to investigate how our algorithmic approach to phonology informs the question of when rule ordering is necessary, especially since the use of the Tolerance Principle allows some degree of robustness to the interaction between rules. That future work may also require adapting the proposed procedure for rule ordering, if cases are identified that fall outside its current purview. Here we provide two interesting cases that exemplify the promise of this direction for future work.

7.3.5.1 Polish Opacity

We demonstrated in § 5.4.3.4 that PLP could handle a case of COUNTERBLEEDING opacity in Polish. Recall that Polish exhibits devoicing in final obstruents, which, as in German and Dutch, manifests in some paradigms like [klup] 'club' ~ [klubɪ] 'clubs' (Sanders, 2003). Some paradigms also exhibit an alternation between low [u] before word-final, underlyingly voiced, oral consonants, and [ɔ] elsewhere, which is usually interpreted as /ɔ/ raising to [u][5] (149; examples from Sanders 2003, ch. 2). Since alternating final obstruents are analyzed as underlyingly voiced, and the environment of raising involves underlying obstruents, the two processes—if productive—must be analyzed as in COUNTERBLEEDING order (149b) with raising preceding devoicing.

(149) a. [dvur] 'mansion' ~ [dvɔri] 'mansions'
[bul] 'ache' ~ [bɔlɛ] 'aches'

b. [bup] 'bean' ~ [bɔbɪ] 'beans'
[lut] 'ice' ~ [lɔdɪ] 'ices'

[5]See Sanders (2003, sec. 2.1.2) for evidence that it is not lowering.

In a brief experimental study, Sanders (2003, sec. 2.3) found that participants did not extend the vowel raising alternation to nonce words in which both alternations would be expected to interact. This suggests that either the interaction between the processes is not productive (i.e. the interaction is synchronically transparent), or that the vowel raising alternation is not productive at all.

It may be the case that either (a) the processes interact rarely enough that the cases of interaction can be lexicalized or (b) the raising alternation may not be pervasive enough to necessitate the learner construct abstract underlying forms for the vowel. In future work, we will investigate both possible explanations in a similar fashion to our study of Dutch (§ 4.4), by using a realistic corpus of child-directed speech.

7.3.5.2 Samala Harmony and Dissimilation

Samala exhibits anticipatory sibilant harmony, exemplified in (150; data from McMullin 2016), in which sibilants agree in anteriority with any sibilant to the right.

(150) /k-su-ʃojin/ → [kʃuʃojin] 'I darken it'
/s-api-tʃʰo-us/ → [sapitsʰolus] 'he has a stroke of good luck'
/s-api-tʃʰo-us-waʃ/ → [ʃapitʃʰoluʃwaʃ] 'he had a stroke of good luck'
/k-su-k'ili-mekeken-ʃ/ → [kʃuk'ilimekeketʃ] 'I straighten myself up'

Samala also has a process by which /s/ dissimilates to [ʃ] to avoid *[st], *[sn], *[sl] structures (151), but harmony overrides this process if both are applicable (151b).

(151) a. /s-tepuʔ/ → [ʃtepuʔ] 'he gambles'
/s-niʔ/ → [ʃniʔ] 'his neck'
/s-lok'in/ → [ʃlok'in] 'he cuts it'
/s-is-tɨʔ/ → [ʃiʃtɨʔ] 'he finds it'
b. /s-iʃ-tiʃi-jep-us/ → [sistisijepus] 'they (dual) show him'
/s-net-us/ → [snetus] 'he does it to him'

As McMullin (2016); De Santo and Graf (2019) discuss, the composition of these processes is not tier strictly local, because predicting the surface form of an underlying /s/ requires that {t, n, l} be excluded from the tier to allow sibilants to be adjacent, while also requiring they be included on the tier to allow for the environment of dissimilation to be visible. Clearly, both requirements cannot be met simultaneously. This provides a proof by negative example that tier-strictly local generalizations are not closed under composition. De Santo and Graf propose an extension to tier strict locality in which the tier projection—originally a context-free function—can be extended to take into account a strictly local context on the input string. This increases the posited expressivity of phonology, which is required if one is interested in characterizing the computational properties

of *grammars*, but is not necessary if one is concerned with individual processes: McMullin (2016, p. 142) demonstrated that the two processes and their interaction can be captured by a constraint ranking of two tier strictly local constraints, and one could accomplish the same with tier strictly local rules by ordering the dissimilation rule before the harmony rule. In chapter § 5, we demonstrated how PLP could achieve some rule orderings and how it may need to be extended to capture more. Future work will investigate whether the model can learn this ordering of rules. If it is possible to learn the ordering, then the this would suggest that the additional computational complexity of the grammar resulting from their interaction is an emergent property, and provides an example of how our approach of directly studying the processes involved in constructing a phonological system complements work on formal-language-theoretic characterizations.

7.3.6 Tonal Phonology

This dissertation has focused on segmental phonology. Another important domain is *tonal phonology*. This is true not only because of coverage of phonological phenomena must include tonal phenomena, but also because tonal phonology plays a critical role in key theoretical issues. For instance, Hyman (2018)'s defense of underlying forms focused on tonal cases. This was no doubt in part to Hyman's area of expertise, but also demonstrates how tonal phenomena center prominently in debates about levels of abstractness. Furthermore, Jardine (2016a) has argued that tonal phonology contains processes, including *tonal plateauing*, that appear to require the ability to capture dependencies that are arbitrarily far away in both directions, in contrast to the non-local processes discussed in Chapter § 3, which involved distant dependencies in only one direction. Clearly then, turning the lens of our algorithmic approach toward tonal phonology has the potential to provide insight into important questions in our understanding of phonology.

7.3.7 Connections to Syntax

Search-and-Copy approaches to non-local phenomena in phonology (Nevins, 2010; Samuels, 2011; Andersson et al., 2020) recognize the similarity to syntactic AGREE (Chomsky, 2001b,a). In both, a recipient searches for a valid donor to specify an unspecified feature value, and take their value from the closest valid donor (we refer readers to Nevins 2010 for extensive discussion). Formal-language-theoretic work has also recognized the relationship between tier-based phonological generalizations and syntactic dependencies, and begun attempts to formalize the relationship (Graf and Shafiei, 2019; Graf and Santo, 2019; Shafiei, 2022).

The notion of closeness differs in syntax and phonology (e.g., c-command vs. string proximity), but this may be a reflection of the data structure over which dependencies are computed, rather than the dependency-inducing operation itself. In our view, the conspicuous similarities between the two

phenomena suggest that the same cognitive mechanism may be at play in both. This view remains somewhat conjectural and needs more investigation, but we view it as a promising line of inquiry, and adopted the terminology AGREE and DISAGREE in chapter § 3 to emphasize this connection.

APPENDIX A

Chapter 4 Appendix

A.1 Theoretical Connections

A.1.1 Relationship to Search and Copy

A number of search-based accounts of vowel harmony have been proposed (Nevins, 2005; Frédéric and Reiss, 2007; Samuels, 2009a,b; Nevins, 2010; Andersson et al., 2020), in which underspecified vowels trigger a linear search for a valid donor from which to copy a feature value for any unspecified feature(s); the search skips invalid donors. Search-and-copy accounts do not provide an account of how the set of segments skipped during the search is learned, which D2L provides via the deletion set. In a tier-projection account, a tier representation is explicitly projected, which renders the relevant dependency *adjacent*. In contrast, search-and-copy operates over an input sequence, skipping the 'non-tier' elements but not projecting an explicit tier. Our reason for treating the tier as a projection is the experimental evidence that adjacency is cognitively prominent (see discussion in § 5.1).

A.1.2 Relationship to Tier-Strict Locality

Since the rules described above (23) involve a tier-local left or right context, they relate to tier-based strictly local functions (Burness et al., 2021) and, because they apply iteratively, they relate specifically to the output-strictly local variant. In these functions, each output segment depends only on previous output segments within a fixed distance k. In our case $k = 2$ McMullin and Hansson (2016), but this could be extended to $k > 2$ by the same process as the strictly-local generalizations from § 5.

A.2 LSTM Details

We use a Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber 1997) as our RNN architecture because it is better-suited for capturing long-distance dependencies than the older models used in phonology (Hare, 1990; Rodd, 1997). We used a Pytorch (Paszke et al., 2019) implementation. Each input sequence is first run through an embedding layer with learned weights, which maps each segment to a real-valued vector representation (embedding) of dimension $d \in \mathbb{N}$. These embeddings are then run through a bidirectional LSTM, which yields a real-valued hidden representation of dimension $h \in \mathbb{N}$ for each embedding. By using a bidirectional LSTM (as opposed to a unidirectional LSTM), the model is capable of capturing dependencies to both the left and the right of an alternating segment, which is important for modeling either leftward- or rightward-spreading processes. Each hidden representation is then mapped to the output alphabet (i.e. the set of possible surface forms for the underlying, input segment) by a fully-connected layer, and a softmax converts the scores to a distribution over the output alphabet. We use a cross-entropy loss, which encourages—via backpropagating the error—the model to assign high probability (≈ 1) to the correct surface form and low probability (≈ 0) to all other possible surface forms. Since we wish to train the model only to predict the surface form of each underlyingly underspecified segments, the cross-entropy loss ignores the model's predictions for underlyingly specified segments. As noted in the text, this makes the problem strictly *easier* for the LSTM than requiring that it predict the *entire* surface sequence. We make this simplification because we think it makes for a more fair comparison to D2L, which also has access to the underlying forms, from which it can derive which segments are alternating.

We train the model on each dataset using the Adam optimizer (Kingma and Ba, 2015) and perform hyperparameter tuning with the Optuna Python package (Akiba et al., 2019). We used the Tree-structured Parzen Estimator (TPE) algorithm, which improves hyperparameter search compared to simpler methods like grid or random search. For hyperparameter tuning, we trained a model with each hyperparameter combination over an 80% sample of the training data, using the remaining 20% as validation data. We carried out the TPE algorithm for 30 iterations, which allows for 30 combinations of hyperparameters to be tested. After these 30 iterations, we used the hyperparameters with the smallest validation loss to train a final model on the entire training data. Hyperparameters were the embedding dimension $d \in \{16, 32, 64, 128, 256, 512\}$, hidden dimension $h \in \{16, 32, 64, 128, 256, 512\}$, learning rate in [0.0001, 0.1], and number of epochs in $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

A.3 Feature Specifications

Table A.1: Features for comparison to Finley (2011).

SEG	cons	voice	son	nas	sib	lab	cor	ant	strid	back	round	hi	low
a	−	+	+							−	−	−	+
b	+	+	−	−	−	+	−	−	−				
d	+	+	−	−	−	−	+	+	−				
e	−	+	+							−	−	−	−
g	+	+	−	−	−	−	−	−	−				
i	−	+	+							−	−	+	−
k	+	−	−	−	−	−	−	−	−				
m	+	+	+	+	−	+	−	−	−				
n	+	+	+	+	−	−	+	+	−				
o	−	+	+							+	+	−	−
p	+	−	−	−	−	+	−	−	−				
s	+	−	−	−	+	−	+	+	+				
t	+	−	−	−	−	−	+	+	−				
u	−	+	+							+	+	+	−
ʃ	+	−	−	−	+	−	+	−	+				

Table A.2: Features for comparison to McMullin and Hansson (2019).

SEG	cons	voice	cont	son	nas	sib	lab	cor	ant	strid	lat	back	hi
b	+	+	−	−	−	−	+	−	−	−	−		
d	+	+	−	−	−	−	−	+	+	−	−		
e	−	+		+								−	−
g	+	+	−	−	−	−	−	−	−	−	−		
i	−	+		+								−	+
k	+	−	−	−	−	−	−	−	−	−	−		
l	+	+	−	+	−	−	−	+	+	−	+		
m	+	+	−	+	+	−	+	−	−	−	−		
n	+	+	−	+	+	−	−	+	+	−	−		
o	−	+		+								+	−
p	+	−	−	−	−	−	+	−	−	−	−		
t	+	−	−	−	−	−	−	+	+	−	−		
u	−	+		+								+	+
ɹ	+	+	−	+	−	−	−	−	−	−	−		

Table A.3: Features for Turkish.

SEG	cons	voice	cont	son	nas	sib	lab	cor	ant	strid	lat	back	round	hi
b	+	+	−	−	−	−	+	−	−	−	−			
d	+	+	−	−	−	−	−	+	+	−	−			
e	−	+		+								−	−	−
f	+	−	+	−	−	−	+	−	−	−	−			
g	+	+	−	−	−	−	−	−	−	−	−			
h	+	−	+	−	−	−	−	−	−	−	−			
i	−	+		+								−	−	+
j	+	+	−	+	−	−	−	−	−	−	−			
k	+	−	−	−	−	−	−	−	−	−	−			
l	+	+	−	+	−	−	−	+	+	−	+			
m	+	+	−	+	+	−	+	−	−	−	−			
n	+	+	−	+	+	−	−	+	+	−	−			
o	−	+		+								+	+	−
p	+	−	−	−	−	−	+	−	−	−	−			
r	+	+	−	+	−	−	−	+	−	−	−			
s	+	−	+	−	−	+	−	+	+	+	−			
t	+	−	−	−	−	−	−	+	+	−	−			
u	−	+		+								+	+	+
v	+	+	+	−	−	−	+	−	−	−	−			
w	+	+	−	+	−	−	+	−	−	−	−			
y	−	+		+								−	+	+
z	+	+	+	−	−	+	−	+	+	+	−			
ø	−	+		+								−	+	−
ɑ	−	+		+								+	−	−
ɯ	−	+		+								+	−	+
ɯ̨	+	+	−	+	−	−	−	−	−	−	−			
ʃ	+	−	+	−	−	+	−	+	−	+	−			
ʒ	+	+	+	−	−	+	−	+	−	+	−			
ʤ	+	+	−	−	−	−	−	+	−	+	−			
ʧ	+	−	−	−	−	−	−	+	−	+	−			

Table A.4: Features for Finnish.

SEG	cons	voice	cont	son	nas	sib	lab	cor	ant	strid	lat	back	round	hi	low
b	+	+	−	−	−	−	+	−	−	−	−				
d	+	+	−	−	−	−	−	+	+	−	−				
e	−	+		+								−	−	−	−
f	+	−	+	−	−	−	+	−	−	−	−				
g	+	+	−	−	−	−	−	−	−	−	−				
h	+	−	+	−	−	−	−	−	−	−	−				
i	−	+		+								−	−	+	−
j	+	+	−	+	−	−	−	−	−	−	−				
k	+	−	−	−	−	−	−	−	−	−	−				
l	+	+	−	+	−	−	−	+	+	−	+				
m	+	+	−	+	+	−	+	−	−	−	−				
n	+	+	−	+	+	−	−	+	+	−	−				
o	−	+		+								+	+	−	−
p	+	−	−	−	−	−	+	−	−	−	−				
r	+	+	−	+	−	−	−	+	−	−	−				
s	+	−	+	−	−	+	−	+	+	+	−				
t	+	−	−	−	−	−	−	+	+	−	−				
u	−	+		+								+	+	+	−
v	+	+	+	−	−	−	+	−	−	−	−				
w	+	+	−	+	−	−	+	−	−	−	−				
y	−	+		+								−	+	+	−
z	+	+	+	−	−	+	−	+	+	+	−				
æ	−	+		+								−	−	−	+
ø	−	+		+								−	+	−	−
ɑ	−	+		+								+	−	−	+

Table A.5: Features for Latin. The consonant features are taken from Cser (2010).

SEG	cons	cor	lab	voice	son	nas	con	lat	high	back	round	hi	low
a	−			+	+					−	−	−	+
b	+	−	+	+	−	−	−	−	−				
d	+	+	−	+	−	−	−	−	−				
e	−			+	+					−	−	−	−
f	+	−	+	−	−	−	+	−	−				
g	+	−	−	+	−	−	−	−	+				
h	+	−	−	+	−	−	+	−	−				
i	−			+	+					−	−	+	−
j	+	+	−	+	+	−	+	−	+				
k	+	−	−	−	−	−	−	−	+				
l	+	+	−	+	+	−	+	+	−				
m	+	−	+	+	+	+	−	−	−				
n	+	+	−	+	+	+	−	−	−				
o	−			+	+					+	+	−	−
p	+	−	+	−	−	−	−	−	−				
r	+	+	−	+	+	−	+	−	−				
s	+	+	−	−	−	−	+	−	−				
t	+	+	−	−	−	−	−	−	−				
u	−			+	+					+	+	+	−
w	+	−	+	+	+	−	+	−	+				

APPENDIX B

Chapter 3 Appendix

B.1 PLP and Strict Locality

We discuss how PLP's generalizations can be characterized in the formal-language-theoretic terms of *strict locality*. We show that the sequences PLP learns are strictly-local definitions, and thus have the interpretation of banning substrings (Heinz, 2018, p. 28) (§ B.1.1). We then discuss how PLP's generalizations describe input-strictly local maps (§ B.1.2).

B.1.1 Strict-Locality of Sequences

Stirctly local stringsets (McNaughton and Papert, 1971) are stringsets whose members 'are distinguished from non-members purely on the basis of their k-factors' (Rogers et al., 2013, p. 98). A k-factor of a string is a length-k substring, and (p. 96) the set of k-factors over an alphabet Σ is $F_k(\Sigma^*) = \{w \in \Sigma^* : |w| \leq k\}$. A *Strictly k-Local Definition* $\mathcal{G}$ is a subset of the k-factors over Σ, i.e., $\mathcal{G} \subseteq F_k(\Sigma^*)$.[1] A definition is a strictly-local definition if it is strictly k-local for some k. We wish to demonstrate that the sequences PLP learns, as defined in (94) and repeated in (152), are strictly-local definitions.

(152) $\mathcal{S} \triangleq \bigcup_{k=1}^{\infty} \{s_1 s_2 ... s_k : s_i \subset \Sigma\}$

Since $\bar{s} \in \mathcal{S}$ is a sequence of *sets* of segments $s_i \subset \Sigma$, we define the *extension*, $E_{\bar{s}}$, of $\bar{s}$ as the set of sequences of *segments* that match $\bar{s}$, as in (153), where $k = |\bar{s}|$.

(153) $E_{\bar{s}} \triangleq \{a_1 a_2 ... a_k : a_i \in s_i \; \forall i \in 1...k\}$

For example, the sequence of two adjacent sibilants (154a) has the extension (154b).

(154) a. $\bar{s}$ = [+sib][+sib]

[1] Rogers et al. (2013) add word-initial '⋊' and word-final '⋉' markers to Σ. We assume the learner's segment inventory already contains symbols for syllable and word boundaries.

b. $E_{\bar{s}} = \{$ss, sz, zs, ʃz, ʒs, ...$\}$

Theorem 1 *The instances $E_{\bar{s}}$ of any $\bar{s} \in \mathcal{S}$ form a Strictly Local definition over the alphabet Σ.*

Proof. For any $a_1a_2...a_k \in E_{\bar{s}}$, each a_i is an element of s_i (i.e., $a_i \in s_i$) by (153) and thus an element of Σ (i.e., $a_i \in s_i \subset \Sigma$) by (152). Thus, every $a_1a_2...a_k \in E_{\bar{s}}$ is a length-k string from Σ^*. It follows that $E_{\bar{s}} \subseteq F_k(\Sigma^*)$ and that, for $k = |\bar{s}|$, $E_{\bar{s}}$ is a Strictly Local Definition. □

B.1.2 Strict-Locality of Generalizations

Chandlee (2014, p. 40) provides formal-language-theoretic and automata-theoretic definitions of *Input Strictly Local* string-to-string functions, which, for input and output alphabets Σ and Γ, have the following interpretation:

Definition 1 (kISL function - Informal) *A function (map) $f : \Sigma^* \rightarrow \Gamma^*$ is Input Strictly Local (ISL) iff $\exists k \in \mathbb{N}$ such that each output symbol $o \in \Gamma$ is determined by a length-k window around its corresponding input symbol.*[2]

Each of PLP's generalizations is interpretable as a rule of the form (155) with a target context (cad) of finite length $|cad|$,[3] and under simultaneous application (cf. § 5.2.3.4).

(155) $a \rightarrow b \,/\, c \_\_ \, d$

Chandlee (2014, p. 41) provides an algorithm for constructing, from any such rule (i.e., with finite target context and under simultaneous application), a Finite State Transducer with the necessary and sufficient automata-theoretic properties of an ISL map. Consequently, if Chandlee's algorithm is a valid *constructive proof*, it follows that each generalization that PLP constructs describes an ISL map. When these are combined into a grammar, it is unknown whether the resulting grammar is also ISL because it is unknown whether ISL maps are closed under composition (Chandlee, 2014, p. 149).

B.2 Differences between PLP and MGL

PLP differs in several ways from the MGL model of Albright and Hayes (2002, 2003). Note that PLP is designed to learn phonology, while MGL was designed for producing English past-tense inflections from verb stems, though it can be extended to other settings.

[2] *Length k includes the corresponding input symbol.*

[3] Under the realistic assumption that input strings are of finite length.

B.2.1 Generalization Strategy

PLP and MGL use different generalization strategies. PLP generalizes as *locally* as possible and MGL generalizes as *conservatively* as possible. As discussed in § 5.1.1 and tested in § 5.4.2, we believe that PLP's generalization strategy is better-supported by studies of human learning.

B.2.2 Number of Rules

Another difference between PLP and MGL is the number of rules they generate. For German syllable-final devoicing at a vocabulary size of 400 (§ 5.4.3), PLP learns a single rule (156).

(156) [+voi,−son] → [−voi] / __]$_\sigma$

In contrast, MGL learns 102 rules for where devoicing should take place and 4138 for where it should not. An example of the former is (157a) and the latter (157b) (both are presented with the extensions of the natural classes for clarity). Rules like (157b) are learned because not every word involves devoicing, and thus MGL needs such rules in order to produce those words (§ B.2.2.1).

(157) a. g → k / {a,e,i,o,u,y,ø,æ,ɔ,ə,ɛ,ɪ,ʊ} __]$_\sigma$ #
⋮

b. ∅ → ∅ / {f,k,p,t,x}]$_\sigma$ #
⋮

B.2.2.1 Production

MGL may produce multiple candidate outputs for an input, because every rule that applies to the input generates a candidate output. The quality of a candidate output is 'the confidence of the best rule that derives it' (Albright and Hayes, 2002, sec. 3.2). We used the candidate with the highest confidence as MGL's prediction. This differs from PLP's production (§ 5.2.3.4), which applies all rules (here just one) in order. This difference is not significant when learning a single phonological process, but it is not straight-forward to use MGL to learn multiple processes simultaneously. For instance, in § 5.4.4, for input /ɪnsɛkt-z/, MGL may have rule(s) for vowel nasalization that produce the candidate *[ĩnsɛktz] and rule(s) for pluralization that produce the candidate *[ɪnsɛkts]. However, MGL does not provide a mechanism to apply *both* rules to produce the correct output [ĩnsɛkts].

B.2.3 Natural Classes

MGL's natural class induction differs from PLP's in two ways. First, MGL does not form natural classes for every part of a rule. For example, the two rules in (158a) will combine to form a third

(158b)—and similarly for (158c) and (158d)—but the rules (158b) and (158d) will not combine to form (158e) because only contexts (not targets) are merged.

(158) a. ə → ə̃ / __ n
ə → ə̃ / __ m
b. ə → ə̃ / __ {n, m}
c. ʌ → ʌ̃ / __ n
ʌ → ʌ̃ / __ m
d. ʌ → ʌ̃ / __ {n, m}
e. {ə, ʌ} → [+nas] / __ {n, m} (formed by PLP but not MGL)

Moreover, when rules are combined, the new rule and the original rules are all retained. In contrast, PLP will construct (158e), and only it will be present in the grammar (§ 5.2.3.1).

Second, when MGL creates natural classes for a set of segments, it retains *all* features shared by those segments, whereas PLP only retains those needed to keep the rule satisfactorily accurate. Thus, for (159a), MGL will construct (159b) while PLP will construct (159c).

(159) a. ə → ə̃ / __ {n, m}
b. ə → ə̃ / __ [+ant, +cons, +lab, +nas, +son, +voi, −back, −cg, −cont, −cor, −delrel, −hi, −lat, −lo, −long, −round, −sg, −syl, −velaric]
c. ə → ə̃ / __ [+nas]

Consequently, PLP will correctly extend ə-nasalization when preceding 'ŋ,' but MGL will need to wait for such an instance in the training data before constructing the full generalization.

APPENDIX C

Chapter 6 Appendix

Table C.1: 100 Training Pairs. The Experimental Group is presented the pairs, the Control Group is presented with only the stems.

(bibu, bibuf), (bəbə, bəbəf), (bədə, bədəf), (bɛbə, bɛbəf), (didu, diduf)
(didə, didəf), (dubə, dubəf), (dudu, duduf), (dudə, dudəf), (dədu, dəduf)
(dədə, dədəf), (dɛdə, dɛdəf), (pibu, pibuf), (pidu, piduf), (pədə, pədəf)
(pɛdu, pɛduf), (pɛdə, pɛdəf), (tibə, tibəf), (tudə, tudəf), (təbu, təbuf)
(təbə, təbəf), (tədu, tədуf), (tɛbu, tɛbuf), (tɛbə, tɛbəf), (tɛdu, tɛduf)
(bipi, bipiʃ), (bipɛ, bipɛʃ), (bupi, bupiʃ), (bəpi, bəpiʃ), (bɛpɛ, bɛpɛʃ)
(bɛtɛ, bɛtɛʃ), (dipi, dipiʃ), (dipɛ, dipɛʃ), (diti, ditiʃ), (dupɛ, dupɛʃ)
(dutɛ, dutɛʃ), (dəpi, dəpiʃ), (dɛtɛ, dɛtɛʃ), (pitɛ, pitɛʃ), (pəti, pətiʃ)
(pɛpɛ, pɛpɛʃ), (tipi, tipiʃ), (tipɛ, tipɛʃ), (titi, titiʃ), (titɛ, titɛʃ)
(tupɛ, tupɛʃ), (tutɛ, tutɛʃ), (tətɛ, tətɛʃ), (tɛpɛ, tɛpɛʃ), (tɛtɛ, tɛtɛʃ)
(bibdə, bibdəf), (bupdə, bupdəf), (butbu, butbuf), (bəpdu, bəpduf), (bətbə, bətbəf)
(dudbu, dudbuf), (dutbu, dutbuf), (dəpdə, dəpdəf), (dɛdbə, dɛdbəf), (dɛtbə, dɛtbəf)
(pipdə, pipdəf), (pudbu, pudbuf), (putbu, putbuf), (pəpdə, pəpdəf), (pətbə, pətbəf)
(tibdu, tibduf), (tibdə, tibdəf), (tidbə, tidbəf), (tudbu, tudbuf), (tupdə, tupdəf)
(tədbu, tədbuf), (təpdə, təpdəf), (tɛbdə, tɛbdəf), (tɛpdu, tɛpduf), (tɛtbu, tɛtbuf)
(bibtɛ, bibtɛʃ), (bidpɛ, bidpɛʃ), (bubtɛ, bubtɛʃ), (bupti, buptiʃ), (bətpɛ, bətpɛʃ)
(bɛbtɛ, bɛbtɛʃ), (bɛpti, bɛptiʃ), (diptɛ, diptɛʃ), (dubtɛ, dubtɛʃ), (duptɛ, duptɛʃ)
(dədpɛ, dədpɛʃ), (dətpi, dətpiʃ), (dɛpti, dɛptiʃ), (pidpi, pidpiʃ), (pudpi, pudpiʃ)
(puptɛ, puptɛʃ), (pəbtɛ, pəbtɛʃ), (pədpi, pədpiʃ), (pətpɛ, pətpɛʃ), (pɛbtɛ, pɛbtɛʃ)
(pɛtpi, pɛtpiʃ), (pɛtpɛ, pɛtpɛʃ), (tidpɛ, tidpɛʃ), (tubtɛ, tubtɛʃ), (tutpɛ, tutpɛʃ)

Table C.2: 52 Test Pairs—the same for the Experimental and Control Groups.

(bipɔ, bipɔf/bipɔʃ), (dupu, dupuf/dupuʃ), (dutu, dutuf/dutuʃ), (dɔpu, dɔpuf/dɔpuʃ)
(dɔpɔ, dɔpɔf/dɔpɔʃ), (dɛpu, dɛpuf/dɛpuʃ), (tipɔ, tipɔf/tipɔʃ), (tɔpu, tɔpuf/tɔpuʃ)
(tɔtu, tɔtuf/tɔtuʃ), (tɛpɔ, tɛpɔf/tɛpɔʃ), (bɔbi, bɔbif/bɔbiʃ), (bɔdɛ, bɔdɛf/bɔdɛʃ)
(bɛdɛ, bɛdɛf/bɛdɛʃ), (dudɛ, dudɛf/dudɛʃ), (pubɛ, pubɛf/pubɛʃ), (pɔdi, pɔdif/pɔdiʃ)
(pɛbɛ, pɛbɛf/pɛbɛʃ), (tidi, tidif/tidiʃ), (tudɛ, tudɛf/tudɛʃ), (tɔbi, tɔbif/tɔbiʃ)
(bubtɔ, bubtɔf/bubtɔʃ), (bɔdpɔ, bɔdpɔf/bɔdpɔʃ), (dɔbtɔ, dɔbtɔf/dɔbtɔʃ), (dɔtpu, dɔtpuf/dɔtpuʃ)
(dɛptɔ, dɛptɔf/dɛptɔʃ), (pɔdpu, pɔdpuf/pɔdpuʃ), (titpu, titpuf/titpuʃ), (tubtu, tubtuf/tubtuʃ)
(tuptu, tuptuf/tuptuʃ), (tɛdpu, tɛdpuf/tɛdpuʃ), (bipdɛ, bipdɛf/bipdɛʃ), (budbi, budbif/budbiʃ)
(bɛpdɛ, bɛpdɛf/bɛpdɛʃ), (dɔbdi, dɔbdif/dɔbdiʃ), (dɛdbɛ, dɛdbɛf/dɛdbɛʃ), (pipdi, pipdif/pipdiʃ)
(pɛdbi, pɛdbif/pɛdbiʃ), (pɛpdɛ, pɛpdɛf/pɛpdɛʃ), (pɛtbi, pɛtbif/pɛtbiʃ), (tidbi, tidbif/tidbiʃ)
(pidɔ, pidɔf/pidɔʃ), (pɔdu, pɔduf/pɔduʃ), (pɛbɔ, pɛbɔf/pɛbɔʃ), (bɛpi, bɛpif/bɛpiʃ)
(putɛ, putɛf/putɛʃ), (pɛtɛ, pɛtɛf/pɛtɛʃ), (budbɔ, budbɔf/budbɔʃ), (ditbu, ditbuf/ditbuʃ)
(pɛtbu, pɛtbuf/pɛtbuʃ), (dɛbti, dɛbtif/dɛbtiʃ), (tɔtpɛ, tɔtpɛf/tɔtpɛʃ), (tɛpti, tɛptif/tɛptiʃ)

BIBLIOGRAPHY

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2623–2631. ACM.

Aksënova, A. (2020). Sigmapie. https://github.com/alenaks/SigmaPie.

Aksu-Koç, A. A. and Slobin, D. I. (1985). Acquisition of Turkish. In Slobin, D. I., editor, *The crosslinguistic study of language acquisition. Volume 1: The data*, pages 839–878. Lawrence Erlbaum, Hillsdale, NJ.

Albright, A. (2002). *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles.

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

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

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

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

Anderson, S. R. (1981). Why phonology isn't "natural". *Linguistic inquiry*, 12(4):493–539.

Anderson, S. R. (2021). *Phonology in the Twentieth Century*. Number 5 in History and Philosophy of the Language Sciences. Language Science Press, Berlin.

Andersson, S., Dolatian, H., and Hao, Y. (2020). Computing vowel harmony: The generative capacity of search & copy. In *Proceedings of the Annual Meetings on Phonology*, volume 7.

Anglin, J. M. (1993). Vocabulary development: A morphological analysis, with commentary by George A. Miller and Pamela C. Wakefield. *Monographs of the Society for Research in Child Development*, pages i–186.

Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological science*, 9(4):321–324.

Baayen, R. H., Chuang, Y.-Y., and Blevins, J. P. (2018). Inflectional morphology with linear mappings. *The mental lexicon*, 13(2):230–268.

Baayen, R. H., Chuang, Y.-Y., Shafaei-Bajestan, E., and Blevins, J. P. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity*, 2019.

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1996). The celex lexical database (cd-rom).

Baer-Henney, D. and van de Vijver, R. (2012). On the role of substance, locality, and amount of exposure in the acquisition of morphophonemic alternations. *Laboratory Phonology*, 3(2):221–249.

Baković, E. (2011). Opacity and ordering. In Goldsmith, J., Riggle, J., and Yu, A. C., editors, *The handbook of phonological theory*. Wiley-Blackwell, Malden, MA, 2nd edition.

Balestriero, R. and Baraniuk, R. G. (2018). A spline theory of deep learning. In *International Conference on Machine Learning*, pages 374–383. PMLR.

Balestriero, R. and Baraniuk, R. G. (2020). Mad max: Affine spline insights into deep learning. *Proceedings of the IEEE*, 109(5):704–727.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Bédécarrats, A., Chen, S., Pearce, K., Cai, D., and Glanzman, D. L. (2018). Rna from trained aplysia can induce an epigenetic engram for long-term sensitization in untrained aplysia. *Eneuro*, 5(3).

Beguš, G. (2016). Post-nasal devoicing and a probabilistic model of phonological typology. *Ms., Harvard University*.

Beguš, G. (2018). *Unnatural phonology: A synchrony-diachrony interface approach*. PhD thesis, Harvard University.

Beguš, G. (2019). Post-nasal devoicing and the blurring process. *Journal of Linguistics*, 55(4):689–753.

Beguš, G. (2020). Generative adversarial phonology: Modeling unsupervised phonetic and phonological learning with neural networks. *Frontiers in Artificial Intelligence*, 3:44.

Beguš, G. (2022). Local and non-local dependency learning and emergence of rule-like representations in speech data by deep convolutional generative adversarial networks. *Computer Speech & Language*, 71:101244.

Behrens, H. (2006). The input–output relationship in first language acquisition. *Language and cognitive processes*, 21(1-3):2–24.

Behrens, S. J. and Blumstein, S. E. (1988). Acoustic characteristics of english voiceless fricatives: A descriptive analysis. *Journal of Phonetics*, 16(3):295–298.

Belth, C. (2023a). A learning-based account of local phonological processes. *Phonology*. In Press.

Belth, C. (2023b). A learning-based account of phonological tiers. Under Review.

Belth, C. (2023c). Towards a learning-based account of underlying forms: A case study in Turkish. *Proceedings of the Society for Computation in Linguistics*.

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The greedy and recursive search for morphological productivity. In *CogSci*.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Bennett, W. G. (2013). *Dissimilation, consonant harmony, and surface correspondence*. Rutgers The State University of New Jersey-New Brunswick.

Berent, I. (2013). The phonological mind. *Trends in cognitive sciences*, 17(7):319–327.

Berent, I., Steriade, D., Lennertz, T., and Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104(3):591–630.

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

Berkson, K., Davis, S., and Strickler, A. (2017). What does incipient /ay/-raising look like?: A response to josef fruehwald. *Language*, 93(3):e181–e191.

Berwick, R. (1985). *The acquisition of syntactic knowledge*. MIT Press, Cambridge, MA.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Best, C. T., McRoberts, G. W., and Sithole, N. M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by english-speaking adults and infants. *Journal of experimental psychology: human perception and performance*, 14(3):345.

Boersma, P. (1997). How we learn variation, optionality, and probability. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, volume 21, pages 43–58. Citeseer.

Boersma, P. et al. (1999). Optimality-theoretic learning in the praat program. In *IFA proceedings*, volume 23, pages 17–35.

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

Boersma, P. and Pater, J. (2008). Convergence properties of a gradual learning algorithm for harmonic grammar.

Booij, G. (1999). The phonology of dutch.

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

Brown, R. (1973). *A first language: The early stages*. Harvard University Press, Cambridge, MA.

Buckler, H. and Fikkert, P. (2016). Dutch and german 3-year-olds' representations of voicing alternations. *Language and Speech*, 59(2):236–265.

Buckley, E. (2000). On the naturalness of unnatural rules. In *Proceedings from the Second Workshop on American Indigenous Languages. UCSB working papers in linguistics*, volume 9, pages 1–14.

Burness, P. and McMullin, K. (2019). Efficient learning of output tier-based strictly 2-local functions. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 78–90, Toronto, Canada. Association for Computational Linguistics.

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

Calamaro, S. and Jarosz, G. (2015). Learning general phonological rules from distributional information: A computational model. *Cognitive science*, 39(3):647–666.

Caplan, S. and Kodner, J. (2018). The acquisition of vowel harmony from simple local statistics. In *CogSci*.

Caprin, C. and Guasti, M. T. (2009). The acquisition of morphosyntax in Italian: A cross-sectional study. *Applied Psycholinguistics*, 30(1):23–52.

Chan, E. (2008). *Structures and distributions in morphology learning*. PhD thesis, University of Pennsylvania.

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

Chandlee, J., Eyraud, R., and Heinz, J. (2014). Learning strictly local subsequential functions. *Transactions of the Association for Computational Linguistics*, 2:491–504.

Chandlee, J., Heinz, J., and Jardine, A. (2018). Input strictly local opaque maps. *Phonology*, 35(2):171–205.

Chang, E. F., Rieger, J. W., Johnson, K., Berger, M. S., Barbaro, N. M., and Knight, R. T. (2010). Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, 13(11):1428–1432.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on information theory*, 2(3):113–124.

Chomsky, N. (2001a). *Beyond explanatory adequacy*. MIT Working Papers in Linguistics, Cambridge, MA.

Chomsky, N. (2001b). Derivation by phase. In Kenstowicz, M., editor, *Ken Hale: A life in language*, pages 1–52. MIT Press, Cambridge, MA.

Chomsky, N. (2005). Three factors in language design. *Linguistic inquiry*, 36(1):1–22.

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

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

Clements, G. N. (1976). *The autosegmental treatment of vowel harmony*. Indiana University Linguistics Club.

Clements, G. N. (1980). *Vowel harmony in nonlinear generative phonology*. Indiana University Linguistics Club Bloomington.

Clements, G. N. (1985). The geometry of phonological features. *Phonology*, 2(1):225–252.

CMU (2014). The carnegie mellon pronuncing dictionary v0.7b. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

Coetzee, A. W. (2009). Learning lexical indexation. *Phonology*, 26(1):109–145.

Coetzee, A. W. and Pater, J. (2011). The place of variation in phonological theory. In Goldsmith, J., Riggle, J., and Yu, A. C., editors, *The handbook of phonological theory*. Wiley-Blackwell, Malden, MA, 2nd edition.

Coetzee, A. W. and Pretorius, R. (2010). Phonetically grounded phonology and sound change: The case of tswana labial plosives. *Journal of Phonetics*, 38(3):404–421.

Cole, D. T. (1955). *An introduction to Tswana grammar*. Longmans, Green & Co.

Çöltekin, Ç. (2010). A freely available morphological analyzer for turkish. In *LREC*, volume 2, pages 19–28.

Çöltekin, Ç. (2014). A set of open source tools for turkish natural language processing. In *LREC*, pages 1079–1086.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009). *Introduction to algorithms*. MIT press.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016). The sigmorphon 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22.

Cotterell, R., Peng, N., and Eisner, J. (2015). Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics*, 3:433–447.

Crick, F., Asanuma, C., et al. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. *Parallel distributed processing*, 2:333–371.

Cser, A. (2010). The -alis/-aris allomorphy revisited. In *Variation and change in morphology: selected papers from the 13th international morphology meeting*, pages 33–52. John Benjamins Publishing.

De Santo, A. and Graf, T. (2019). Structure sensitive tier projection: Applications and formal properties. In *International Conference on Formal Grammar*, pages 35–50. Springer.

Deen, K. U. (2005). *The acquisition of Swahili*. John Benjamins Publishing, Amsterdam.

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of english. *Language and Speech*, 49(2):137–173.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Dewey, J. (1916). *Democracy and education: An introduction to the philosophy of education*. Macmillan.

Dobrovolsky, M. (1982). Some thoughts on turkish voicing assimilation. In *Calgary Working Papers in Linguistics*, volume 7, pages 1–6.

Eberhard, D. M., Simons, G. F., and Fennig, C. D. (2023). Ethnologue: Languages of the world.

Ellis, K., Albright, A., Solar-Lezama, A., Tenenbaum, J. B., and O'Donnell, T. J. (2022). Synthesizing theories of human language with bayesian program induction. *Nature communications*, 13(1):5024.

Emond, E. and Shi, R. (2021). Infants' rule generalization is governed by the Tolerance Principle. In Dionne, D. and Vidal Covas, L.-A., editors, *Proceedings of the 45nd annual Boston University Conference on Language Development*, pages 191–204.

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

Ettlinger, M. (2008). *Input-driven opacity*. University of California, Berkeley.

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

Fikkert, P. (1994). *On the acquisition of prosodic structure*. [Sl: sn].

Finley, S. (2011). The privileged status of locality in consonant harmony. *Journal of memory and language*, 65(1):74–83.

Finley, S. (2015). Learning nonadjacent dependencies in phonology: Transparent vowels in vowel harmony. *Language*, 91(1):48.

Finley, S. (2017). Locality and harmony: Perspectives from artificial grammar learning. *Language and Linguistics Compass*, 11(1):e12233.

Finley, S. and Badecker, W. (2009). Artificial language learning and feature-based generalization. *Journal of memory and language*, 61(3):423–437.

Fiser, J. and Aslin, R. N. (2002). Statistical learning of higher-order temporal structure from visual shape sequences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):458.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Frédéric, M. and Reiss, C. (2007). Computing long-distance dependencies in vowel harmony. *Biolinguistics*, 1:28–48.

Freitas, M. J. (2003). The acquisition of onset clusters in european portuguese.

Gafos, A. I. (2014). *The articulatory basis of locality in phonology*. Routledge.

Gallistel, C. R. (2000). The replacement of general-purpose learning models with adaptively specialized learning modules. *The new cognitive neurosciences*, pages 1179–1191.

Gallistel, C. R. and King, A. P. (2011). *Memory and the computational brain: Why cognitive science will transform neuroscience*, volume 6. John Wiley & Sons.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *International conference on machine learning*, pages 1243–1252. PMLR.

Gershman, S. J., Balbi, P. E., Gallistel, C. R., and Gunawardena, J. (2021). Reconsidering the evidence for learning in single cells. *Elife*, 10:e61907.

Gildea, D. and Jurafsky, D. (1996). Learning bias and phonological-rule induction. *Computational Linguistics*, 22(4):497–530.

Gnanadesikan, A. (2004). Markedness and faithfulness constraints in child phonology. *Constraints in phonological acquisition*, pages 73–108.

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

Goldsmith, J. (1976). *Autosegmental phonology*. PhD thesis, Massachusetts Institute of Technology.

Goldsmith, J. (1985). Vowel harmony in khalkha mongolian, yaka, finnish and hungarian. *Phonology Yearbook*, 2(1):253–275.

Goldsmith, J. and Laks, B. (2006). Generative phonology: its origins, its principles, and its successors. *Waugh, Linda, E. Joseph, John, E. The Cambridge History of Linguistics*.

Goldsmith, J. and Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3):859–896.

Goldsmith, J. and Xanthos, A. (2009). Learning phonological categories. *Language*, 85(1):4–38.

Goldwater, S., Johnson, M., Spenader, J., Eriksson, A., and Dahl, Ö. (2003). Learning ot constraint rankings using a maximum entropy model. In *Proceedings of the Stockholm workshop on variation within Optimality Theory*, volume 111, page 120.

Gómez, R. and Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy*, 7(2):183–206.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13(5):431–436.

Gould, S. J. (1989). *Wonderful life: the Burgess Shale and the nature of history*. WW Norton & Company.

Gould, S. J. (1999). *Questioning the Millennium: A Rationalist's Guide to a Precisely Arbitrary Countdown (Revised Edition)*. Crown.

Gould, S. J. (2002). *The structure of evolutionary theory*. Harvard university press.

Gouskova, M. and Gallagher, G. (2020). Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory*, 38(1):77–116.

Graf, T. and Santo, A. D. (2019). Sensing tree automata as a model of syntactic dependencies. In *Proceedings of the 16th Meeting on the Mathematics of Language*.

Graf, T. and Shafiei, N. (2019). C-command dependencies as TSL string constraints. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, pages 205–215.

Graves, A., Wayne, G., and Danihelka, I. (2014). Neural turing machines. *CoRR*, abs/1410.5401.

Grijzenhout, J. and Joppen, S. (1998). First steps in the acquisition of german phonology: A case study. Seminar für Allgemeine Sprachwissenschaft, Heinrich-Heine-Univ.

Grijzenhout, J. and Joppen-Hellwig, S. (2002). The lack of onsets in german child phonology. *The Process of Language Acquisition. Frankfurt am Main: Peter Lang Verlag*, pages 319–339.

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

Halle, M. (1978). *Knowledge unlearned and untaught: What speakers know about the sounds of their language*. na.

Hare, M. (1990). The role of similarity in hungarian vowel harmony: a connectionist account. *Connection Science*, 2(1-2):123–150.

Hart, B. and Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H Brookes Publishing, Baltimore, MD.

Hayes, B. (2004). Phonological acquisition in optimality theory: the early stages. *Constraints in phonological acquisition*, pages 158–203.

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

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

Hayes, B. and Stivers, T. (2000). Postnasal voicing. *Ms., UCLA*.

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

Hayward, R. J. (1990). Notes on the Aari language. *Omotic language studies*, pages 425–493.

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

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

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

Herdan, G. (1960). *Type-token mathematics*, volume 4. Mouton.

Hewitt, J. and Manning, C. D. (2019). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Hosseini, E. A., Schrimpf, M. A., Zhang, Y., Bowman, S., Zaslavsky, N., and Fedorenko, E. (2022). Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *bioRxiv*, pages 2022–10.

Howard, I. (1972). *A directional theory of rule application in phonology.* PhD thesis, Massachusetts Institute of Technology.

Hua, W., Jardine, A., and Dai, H. (2020). Learning underlying representations and input-strictly-local functions. In *Proceedings of the 37th West Coast Conference on Formal Linguistics*.

Hyman, L. M. (2018). Why underlying representations? *Journal of Linguistics*, 54(3):591–610.

Jakobson, R. and Halle, M. (1956). *Fundamentals of language*. De Gruyter Mouton.

Jardine, A. (2016a). Computationally, tone is different. *Phonology*, 33(2):247–283.

Jardine, A. (2016b). Learning tiers for long-distance phonotactics. In *Proceedings of the 6th conference on Generative Approaches to Language Acquisition North America (GALANA)*, pages 60–72.

Jardine, A., Chandlee, J., Eyraud, R., and Heinz, J. (2014). Very efficient learning of structured classes of subsequential functions from positive data. In *International Conference on Grammatical Inference*, pages 94–108. PMLR.

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

Jardine, A. and McMullin, K. (2017). Efficient learning of tier-based strictly k-local languages. In Drewes, F., Martín-Vide, C., and Truthe, B., editors, *Language and Automata Theory and Applications*, pages 64–76. Springer.

Jarosz, G. (2019). Computational modeling of phonological learning. *Annual Review of Linguistics*.

Joanisse, M. F. and McClelland, J. L. (2015). Connectionist perspectives on language learning, representation and processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, 6(3):235–247.

Johansson, F., Jirenhed, D.-A., Rasmussen, A., Zucca, R., and Hesslow, G. (2014). Memory trace and timing mechanism localized to cerebellar purkinje cells. *Proceedings of the National Academy of Sciences*, 111(41):14930–14934.

Johnsen, S. S. (2012). A diachronic account of phonological unnaturalness. *Phonology*, 29(3):505–531.

Johnson, C. D. (1972). *Formal aspects of phonological description*, volume 3. The Hague.

Johnson, M. (1984). A discovery procedure for certain phonological rules. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*, pages 344–347.

Joos, M. (1942). A phonological dilemma in canadian english. *Language*, pages 141–144.

Jusczyk, P. W., Smolensky, P., and Allocco, T. (2002). How english-learning infants respond to markedness and faithfulness constraints. *Language Acquisition*, 10(1):31–73.

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

Kaiser, L. and Sutskever, I. (2016). Neural gpus learn algorithms. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Kaplan, R. M. and Kay, M. (1994). Regular models of phonological rule systems. *Computational linguistics*, 20(3):331–378.

Kenstowicz, M. (1994). *Phonology in generative grammar*. Blackwell, Malden, MA.

Kenstowicz, M. and Kisseberth, C. (1977). *Topics in phonological theory*. Academic Press, New York.

Kenstowicz, M. and Kisseberth, C. (1979). *Generative phonology: Description and theory*. Academic Press, San Diego.

Kerkhoff, A. O. (2007). *Acquisition of morpho-phonology: The Dutch voicing alternation*. PhD thesis, LOT.

Kharitonov, E., Lee, A., Polyak, A., Adi, Y., Copet, J., Lakhotia, K., Nguyen, T. A., Riviere, M., Mohamed, A., Dupoux, E., and Hsu, W.-N. (2022). Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681, Dublin, Ireland. Association for Computational Linguistics.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kiparsky, P. (1968). *How abstract is phonology?* Indiana University Linguistics Club.

Kiparsky, P. (1973). *Abstractness, opacity and global rules*. Indiana University Linguistics Club.

Kirov, C. and Cotterell, R. (2018). Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665.

Kodner, J. (2022). Computational models of morphological learning. In *Oxford Research Encyclopedia of Linguistics*.

Korn, D. (1969). Types of labial vowel harmony in the Turkic languages. *Anthropological Linguistics*, 11(3):98–106.

Kornfilt, J. (2013). *Turkish*. Routledge.

Koulaguina, E. and Shi, R. (2019a). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, 26(4):416–435.

Koulaguina, E. and Shi, R. (2019b). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, 26(4):416–435.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

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

Labov, W. (1969). Contraction, deletion, and inherent variability of the English copula. *Language*, 45(4):715–762.

Ladányi, E., Kovács, Á. M., and Gervain, J. (2020). How 15-month-old infants process morphologically complex forms in an agglutinative language? *Infancy*, 25(2):190–204.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al. (2021). On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Lee, J. L., Ashby, L. F., Garza, M. E., Lee-Sikka, Y., Miller, S., Wong, A., McCarthy, A. D., and Gorman, K. (2020). Massively multilingual pronunciation modeling with WikiPron. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4223–4228, Marseille, France. European Language Resources Association.

Legendre, G., Miyata, Y., and Smolensky, P. (1990). Harmonic grammar—a formal multi-level connectionist theory of linguistic well-formedness: Theoretical foundations. In *Proceedings of the twelfth annual conference of the Cognitive Science Society*, pages 884–891. Citeseer.

Levelt, W. J. (2013). *A history of psycholinguistics: The pre-Chomskyan era*. Oxford University Press.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, 74(6):431.

Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic inquiry*, 8(2):249–336.

Lignos, C. and Yang, C. (2016). Morphology and language acquisition. In Hippisley, Andrew R. abd Stump, G., editor, *The Cambridge handbook of Morphology*, chapter 28, pages 765–791. Cambridge University Press, Cambridge.

Lindblad, V. M. (1990). *Neutralization in Uyghur*. University of Washington.

Liu, P., Chang, S., Huang, X., Tang, J., and Cheung, J. C. K. (2019). Contextualized non-local neural networks for sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6762–6769.

Locke, J. L. (1983). *Phonological acquisition and change*. Academic Press.

MacWhinney, B. (1978). *The acquisition of morphophonology*. Monographs of the Society for Research in Child Development. University of Chicago Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. transcription format and programs*, volume 1. Psychology Press.

Marquis, A. and Shi, R. (2012). Initial morphological learning in preverbal infants. *Cognition*, 122(1):61–66.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

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

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

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic inquiry*, 12(3):373–418.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica*, 45(2-4):84–108.

McCarthy, J. J. (2007). *Derivations and levels of representation*, pages 99–118. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1987). *Parallel Distributed Processing, Volume 2: Explorations in the Microstructure of Cognition: Psychological and Biological Models*, volume 2. MIT press.

McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. (1986). The appeal of parallel distributed processing. *MIT Press, Cambridge MA*, 3:44.

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

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133.

McCurdy, K., Goldwater, S., and Lopez, A. (2020). Inflecting when there's no majority: Limitations of encoder-decoder neural networks as cognitive models for German plurals. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1745–1756. Association for Computational Linguistics.

McMullin, K. and Hansson, G. Ó. (2016). Long-distance phonotactics as tier-based strictly 2-local languages. In *Proceedings of the annual meetings on phonology*, volume 2.

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

McMullin, K. J. (2016). *Tier-based locality in long-distance phonotactics: learnability and typology*. PhD thesis, University of British Columbia.

McNaughton, R. and Papert, S. A. (1971). *Counter-Free Automata (MIT research monograph no. 65)*. The MIT Press.

Mesgarani, N., Cheung, C., Johnson, K., and Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010.

Mintz, T. H. (2013). The segmentation of sub-lexical morphemes in english-learning 15-month-olds. *Frontiers in psychology*, 4:24.

Mortensen, D. R., Littell, P., Bharadwaj, A., Goyal, K., Dyer, C., and Levin, L. (2016). PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Narayan, C. R., Werker, J. F., and Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental science*, 13(3):407–420.

Nastase, V., Mihalcea, R., and Radev, D. R. (2015). A survey of graphs in natural language processing. *Natural Language Engineering*, 21(5):665–698.

Nevins, A. (2005). *Conditions on (dis) harmony*. PhD thesis, Massachusetts Institute of Technology.

Nevins, A. (2010). *Locality in vowel harmony*, volume 55. Mit Press.

Newport, E. L. and Aslin, R. N. (2004). Learning at a distance i. statistical learning of non-adjacent dependencies. *Cognitive psychology*, 48(2):127–162.

Oflazer, K. (1994). Two-level description of turkish morphology. *Literary and linguistic computing*, 9(2):137–148.

O'Hara, C. (2017). How abstract is more abstract? learning abstract underlying representations. *Phonology*, 34(2):325–345.

Oncina, J., García, P., and Vidal, E. (1993). Learning subsequential transducers for pattern recognition interpretation tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(5):448–458.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pater, J. (1999). Austronesian nasal substitution and other nc effects. *The prosody-morphology interface*, pages 310–343.

Payne, S. (2022). *When collisions are a good thing: the acquisition of morphological marking*. Bachelor's thesis, University of Pennsylvania.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., and Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, 101(3):B31–B41.

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

Podlipniak, P. (2017). The role of the baldwin effect in the evolution of human musicality. *Frontiers in neuroscience*, 11.

Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., and Dupoux, E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. In *Proceedings of Interspeech*.

Prince, A. and Smolensky, P. (1993). Optimality Theory: Constraint interaction in generative grammar. Technical report.

Rasin, E., Berger, I., Lan, N., and Katzir, R. (2018). Learning phonological optionality and opacity from distributional evidence. In *Proceedings of NELS*, volume 48, pages 269–282.

Reddy, S., Chen, D., and Manning, C. D. (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Reitmaier, T., Wallington, E., Kalarikalayil Raju, D., Klejch, O., Pearson, J., Jones, M., Bell, P., and Robinson, S. (2022). Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Rescorla, M. (2020). The Computational Theory of Mind. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2020 edition.

Richter, C. (2018). Learning allophones: What input is necessary. In *Proceedings of the 42nd annual Boston University Conference on Language Development.* Cascadilla Press.

Richter, C. (2021). *Alternation-Sensitive Phoneme Learning: Implications For Children's Development And Language Change*. PhD thesis, University of Pennsylvania.

Ringe, D. and Eska, J. F. (2013). *Historical linguistics: Toward a twenty-first century reintegration*. Cambridge University Press.

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

Rodd, J. (1997). Recurrent neural-network learning of phonological regularities in Turkish. In *Computational Natural Language Learning (CoNLL)*.

Rogers, J., Heinz, J., Fero, M., Hurst, J., Lambert, D., and Wibel, S. (2013). Cognitive and subregular complexity. In *Formal grammar*, pages 90–108. Springer.

Rose, S. and Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, pages 475–531.

Rosenthall, S. (1989). The phonology of nasal-obstruent sequences.

Rossion, B. and Pourtois, G. (2004). Revisiting snodgrass and vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, 33(2):217–236.

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In McClelland, J. L., Rumelhart, D. E., and the PDP Research Group, editors, *Parallel distributed processing: Explorations into the microstructure of cognition. Volume 2: Psychological and biological models*, pages 216–271. MIT Press, Cambridge, MA.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., and Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52.

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., and Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological science*, 8(2):101–105.

Samuels, B. D. (2009a). Structure & specification in harmony. In *North East Linguistics Society (NELS)*.

Samuels, B. D. (2009b). *The structure of phonological theory*. PhD thesis, Harvard University.

Samuels, B. D. (2011). *Phonological architecture: A biolinguistic perspective*, volume 2. Oxford University Press.

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

Santelmann, L. M. and Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, 69(2):105–134.

Saussure, F. d. (1916). *Cours de linguistique générale*. Éditions Payot & Rivages (Wade Baskin 1959 translation), Paris.

Schaefer, R. P. (1982). A strength hierarchy for a morphophonemic process in Tswana. *Studies in African Linguistics*, 13(2):147–176.

Schmid, H. (1999). Improvements in part-of-speech tagging with an application to german. *Natural language processing using very large corpora*, pages 13–25.

Schmid, H. (2013). Probabilistic part-ofispeech tagging using decision trees. In *New methods in language processing*, page 154.

Schuler, K. D., Yang, C., and Newport, E. L. (2016). Testing the tolerance principle: Children form productive rules when it is more computationally efficient to do so. In *CogSci*, volume 38, pages 2321–2326.

Seidl, A. and Buckley, E. (2005). On the learning of arbitrary phonological rules. *Language Learning and Development*, 1(3–4):289–316.

Shafiei, N. (2022). *Computational Locality and Domain of Syntactic Long-distance Dependencies*. PhD thesis, State University of New York at Stony Brook.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Siegelmann, H. T. and Sontag, E. D. (1992). On the computational power of neural nets. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, page 440–449, New York, NY, USA. Association for Computing Machinery.

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

Smith, C., O'Hara, C., Rosen, E., and Smolensky, P. (2021). Emergent gestural scores in a recurrent neural network model of vowel harmony. *Proceedings of the Society for Computation in Linguistics*, 4(1):61–70.

Smith, D. A., Rydberg-Cox, J. A., and Crane, G. R. (2000). The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25.

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

Smolensky, P. (1996). The initial state and 'richness of the base' in optimality theory. *Rutgers Optimality Archive*, 293.

Smolensky, P. and Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar (Cognitive architecture), Vol. 1*. MIT press.

Snodgrass, J. G. and Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of experimental psychology: Human learning and memory*, 6(2):174.

Suomi, K., Toivanen, J., and Ylitalo, R. (2008). *Finnish Sound Structure: Phonetics, Phonology, Phonotactics and Prosody*, volume 9 of *Studia Humaniora Ouluensia*. University of Oulu.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

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

Tesar, B. (2013). *Output-Driven Phonology: Theory and Learning*. Cambridge University Press, Cambridge.

Tesar, B. and Smolensky, P. (1998). Learnability in optimality theory. *Linguistic Inquiry*, 29(2):229–268.

Turing, A. M. (1937). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, s2-42(1):230–265.

Van de Vijver, R. and Baer-Henney, D. (2014). Developing biases. *Frontiers in psychology*, 5:634.

Van der Hulst, H. (2016). Vowel harmony. In *Oxford Research Encyclopedia of Linguistics*.

Van Kampen, J. (2009). The non-biological evolution of grammar: Wh-question formation in germanic. *Biolinguistics*, 3(2-3):154–185.

Van Kampen, N. J. (1994). The learnability of the left branch condition. *Linguistics in the Netherlands*, 11:83–94.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vaux, B. (2008). Why the phonological component must be serial and rule-based. In Vaux, B. and Nevins, A., editors, *Rules, constraints, and phonological phenomena*, pages 20–61. Oxford University Press, Oxford.

Von Neumann, J. (1945). First Draft of a Report on the EDVAC. Technical report.

Warstadt, A., Choshen, L., Mueller, A., Williams, A., Wilcox, E., and Zhuang, C. (2023). Call for papers–the babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63.

White, J., Kager, R., Linzen, T., Markopoulos, G., Martin, A., Nevins, A., Peperkamp, S., Polgárdi, K., Topintzi, N., and van De Vijver, R. (2018). Preference for locality is affected by the prefix/suffix asymmetry: Evidence from artificial language learning. In *NELS*, pages 207–220.

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

Wiese, R. (1996). The phonology of german.

Wu, L., Chen, Y., Shen, K., Guo, X., Gao, H., Li, S., Pei, J., and Long, B. (2021). Graph neural networks for natural language processing: A survey. *arXiv preprint arXiv:2106.06090*.

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

Yang, C., Crain, S., Berwick, R. C., Chomsky, N., and Bolhuis, J. J. (2017). The growth of language: Universal grammar, experience, and principles of computation. *Neuroscience & Biobehavioral Reviews*, 81:103–119.

Zamuner, T. S., Kerkhoff, A., and Fikkert, J. (2006). Acquisition of voicing neutralization and alternations in dutch. In *Proceedings of the 30th annual Boston University Conference on Language Development*, pages 701–712.

Zamuner, T. S., Kerkhoff, A., and Fikkert, P. (2012). Phonotactics and morphophonology in early child language: Evidence from dutch. *Applied psycholinguistics*, 33(3):481–499.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley, Cambridge, MA.