How Joint Inference of the Lexicon and Phonology Affects the Learnability of Process Interactions

by

Christopher Yang

B.A., Linguistics, Specialization in Computing
University of California, Los Angeles (2017)

Submitted to the Department of Linguistics & Philosophy
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2023

Author .....
Department of Linguistics & Philosophy
September 8, 2023

Certified by .....
Adam Albright
Professor of Linguistics
Thesis Supervisor

Certified by .....
Naomi H. Feldman
Associate Professor of Lingustics and UMIACS, UMD
Thesis Supervisor

Accepted by .....
Daniel Fox
Anshen-Chomsky Professor of Language & Thought
Department Head

How Joint Inference of the Lexicon and Phonology Affects the Learnability of Process Interactions

by

Christopher Yang

Submitted to the Department of Linguistics & Philosophy
on September 8, 2023, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Contemporary phonological research has increasingly become interested in exploring the topic of learnability through the use of computational models. However, many of the proposed models lack one or more of the following properties. ① Many models do not consider the effect of the lexicon at all on performance, and those that do fail to consider the effect contextual allomorphy has on performance. ② Many models characterize learnability in terms of the algorithmic implementation of search, rather than a more principled relationship between the data and the hypothesis space. These properties are critically relevant when it comes to the learnability of process interactions. The experimental literature has demonstrated that artificial languages exhibiting patterns generated from certain process interactions are more likely to be successfully reproduced and generalized by participants than others (Ettlinger 2008; Kim 2012; Brooks, Pajak, & Baković 2013; Prickett 2019). The historical literature has likewise noted that surface patterns generated from particular process interactions are more likely to change in systematic ways than others, including lexicalization, in which an alternation is encoded into the lexicon instead of the phonology, and reanalysis, in which a surface generalization is lost or changed entirely (Kiparsky 1968, 1971). Each of these hypotheses make different predictions when generating forms not seen during training. In this dissertation, I make the following contributions. ① I propose a novel noisy-channel model of morphophonological learning. This model jointly infers a weighted space of consistent and nearly consistent lexicons and grammars from labelled, unparsed surface data. Predictions are generated given the entirety of the inferred weighted space. ② I compare the predictions of the model to the results two artificial language learning experiments, which, despite involving the same underlying processes, produced contradictory results. I show that the model is able to achieve the results of both experiments under a unified account: the generalizability of a pattern is determined by the number of hypotheses compatible or nearly compatible with that pattern.

Thesis Supervisor: Adam Albright
Title: Professor of Linguistics

Thesis Supervisor: Naomi H. Feldman
Title: Associate Professor of Lingustics and UMIACS, UMD

Acknowledgments

The sheer number of people who have contributed to both the development and writing of this dissertation as well as my general well-being during my time in the program is enormous. I am not sure if I will be able to successfully put into words my appreciation to all those who have positively affected my life these past 6 years, but I will do my best.

First, I want to thank the members of my committee, Adam Albright, Naomi Feldman, Michael Kenstowicz, and Donca Steriade. Each and every one of them has been instrumental to the development, writing, and re-writing of this dissertation. I want to thank in particular both Donca and Michael for their help on the lattermost aspect, as without their feedback, the dissertation would have been much less clear and riddled in typos, especially regarding issues with long-distance agreement.

Adam has been my de facto advisor from very early on in the program, and has constantly (but gently) pushed me to strive harder both as a researcher and as a writer. I thank him for always reminding me to think of the bigger picture, course-correcting me whenever I started straying too far away from my original questions, and calling me out (again, gently) when what I am saying did not make sense. If it was not for him, all of my work would be filled with red-herrings. During lockdown, Adam made sure that everyone was keeping their sanity by organizing morning meet-ups, which I am also grateful for. His patience and words of encouragement have been integral to my surviving and completing of this program.

Naomi came to MIT as a visiting professor in my first year, and her presence has shaped my research interests and goals from the very beginning. She was the one who introduced me to the wonders of computational modeling, and who taught me almost everything I know about Bayesian statistics, modeling, and their applications to phonology. I would not be here writing these acknowledgements had it not been for her. Despite me not being one of her official advisees, she has provided more than I could have ever wanted in an advisor, meeting with me weekly to give helpful feedback on my work, providing words of encouragement and wisdom when I was struggling with a challenging problem, and checking in on me whenever I fell off the radar.

The COVID-19 lockdown did a number on my mental health, as I am sure it did to many others. I want to thank MIT Mental Health and Counseling for connecting me to the proper resources to manage my stress, depression, and anxiety. I also want to thank Yiling Zhang for helping me navigate through these darker times and getting me back to a state where I could start functioning again.

I have had the chance to meet so many incredible people in this department. I want

to thank every person in my `ling-17` cohort: Daniel Asherov, Tatiana Bondarenko, Sherry Yong Chen, Cater Fulang Chen, Boer Fu, Filipe Hisao Kobayashi, Vincent Rouillard, Dóra Kata Takács, and Stanislao Zompì. I also thank the various gym and climbing buddies that I have made over the years: Rafael Abramovitz, Daniel Asherov, Christopher Baron, Enrico Flor, Filipe Hisao Kobayashi, Julian Kirkeby Lysvik, Maša Močnik, Jocelyn Wang, Silvan Wittwer, and Frank Staniszewski. Of my departmental seniors, I want to specifically give a shout-out to Cora Lesure for introducing me to my new addiction of knitting, and Erin Olson for being an incredible role-model for me as a striving phonologist. She is also an amazing artist! I also want to give a shout-out to my departmental juniors Hyun Ji Yoo for her wittiness and kindness, and for letting me beat her in our last game of *Dominion* ©, as well as Agnes Ruyue Bi, referred to as "Ag-Ag" only by me. Other current and former members of the department that I wish to thank are Suzana Fong, Verena Hehl, Keely New, Giovanni Roversi, and Abdul-Razak Suleman.

Filipe Hisao Kobayashi has played a variety of different roles in my life during my time here, from being a cohortmate to roommate, a gym-buddy to climbing-buddy, and a colleague to a close friend. He has always been a grounding voice of reason throughout my 6 years here, and a consistent presence particularly during the dissertation writing season. Thank you for introducing me to the world of Brahms, and for getting me so, *so* many donut holes over the course of the past several months. My stomach thanks you.

Daniel Asherov has been one of my closest friends from the beginning of the program. He was always a reliable source for advice, encouragement, and validation, and I never felt shy with reaching out to him and asking for help when I needed it. Thank you for being someone who was always willing to lend an open ear, an individual who inspires me to want to become a better person, and a pal with whom I can watch movies that no-one else wants to see.

Beyond the department, I have had the opportunity to meet and befriend some amazing people. These include former roommates-become-friends Mary Kate Dornon, Maya Ludtke, Matthew McWeeney, and Sydney Stewart, and former acquaintances-become-friends Rebecca Chen, Grace Chow, Erin Chow, and Shaye Firmin. I will never get any better at Nerts, but it is still fun to play.

When I first met Sara Shin in the fall of 2019, I did not expect a full-blown pandemic to take place, resulting in a lockdown lasting several years, nor did I expect to build one of the most meaningful relationships that I have ever been in, yet it seems both have happened. Thank you for putting up with me for the past four years. Your presence and food are always a source of immense comfort, and not something I will ever take for granted.

Back in my hometown of San Jose, California, I have a few more people I want to thank.

I wish to thank my long-time friends Timothy Ho, Chaitanya Mittal, Krishnan Srinivasan, Austin Wang, and Roger Volden for always making me feel at home, no matter how long it has been since I visited. Lastly, I wish to thank my family. To my mom and dad for building a home in a foreign country while raising two rowdy children, and to my sister for constantly checking in on me while I have been living here in Massachusetts, even when I would ignore her texts for weeks. I do not know how you all put up with me for almost thirty years. I love you.

Contents

List of Figures

List of Tables

Chapter 1

Process Interactions, the Lexicon, and the Lexicophonological Space

A fundamental question in linguistics is uncovering how an individual learns the underlying structure of their language given a distribution of positive, unparsed surface data. In the realm of phonological learning, the underlying structures of interest correspond to the LEXICON, or the distribution of underlying forms, henceforth URs, associated with specific meanings, and the PHONOLOGICAL GRAMMAR, or the function that maps URs into their observed counterparts, henceforth ORs.[1] An ever-growing body of literature has focused on the topic of LEARNABILITY. Here, the goal is to investigate whether some phonological patterns are easier for learners to reproduce and generalize than others, and, if so, why. The notion of learnability has been increasingly explored as an explanation for asymmetries observed in the TYPOLOGY. Under this proposal, phonological patterns that are more commonly observed amongst the world's languages emerge due to two factors: ① the pattern is more likely to PERSIST over generations of speakers, and ② the pattern is more likely to be INNOVATED from a language originally lacking it.

One area in which this explanation is commonly invoked in phonology is with respect to the topic of PROCESS INTERACTIONS. Here, the application of one phonological process potentially aids or interferes with another process's ability to apply. For example, in rule terms, consider the processes in (1).

(1) a. DELETION: $V \rightarrow \varnothing$ / _V
b. PALATALIZATION: $t \rightarrow t\int$ / _i

[1] I label outputs as ORs rather than SRs due to a distinction made by my model that does not align with the typical notion of SRs. I will elaborate on this contrast in Chapter 2.

Given the UR /ikit-a-i/, the application of the deletion process creates the environment for which the palatalization process can then apply, i.e. /ikit-a-i/ $\xrightarrow{\text{Del.}}$ [ikiti] $\xrightarrow{\text{Pal.}}$ [ikitʃi]. In contrast, given the underlying form /ikit-i-a/, the application of the deletion process eliminates the environment for which the palatalization process could have applied, i.e. /ikit-i-a/ $\xrightarrow{\text{Del.}}$ [ikita].

In the empirical domain, both the historical and experimental literature have made the claim that patterns generated by some process interactions are more likely to persist and be reproduced by learners of the language than others. The directionality of the asymmetry, however, varies wildly from case to case.

Most contemporary research assessing the sources driving the observed asymmetries in reproduceability utilize COMPUTATIONAL MODELS. Research in this area has primarily been interested in exploring how these differences can be attributed to differences in GRAMMATICAL LEARNABILITY, as described by Heinz (2011) in (2).

(2) "Given a finite set of underlying forms paired with surface forms generated by a particular phonological grammar, what phonological grammar generated them?"
Or, *do some patterns persist due to how easy is it for the learner to recover the underlying grammar than others?*

For example, given the UR /ikit-a-i/ and the OR [ikitʃi], what is the underlying grammar that produced this mapping? How easy is it for the model to recover this grammar compared to one in which the OR is [ikiti]?

While these models have seen success at capturing certain observed empirical asymmetries reported in the literature, many of these computational models omit an integral part of phonological learning: the joint acquisition of the lexicon along with the grammar. This factor is of particular relevance with respect to the topic of process interactions as distributions that appear to have been generated from certain processes interacting in a particular way have been demonstrated to instead emerge from the interaction between the lexicon and the grammar. For example, certain process interactions have been demonstrated to undergo LEXICALIZATION, where an observed phonological alternation becomes unproductive and is instead encoded into the lexicon instead (cf. alternative analyses in which specific process interactions can never be rendered phonologically, Sanders 2003). The lexicon therefore allows for competing hypotheses that, despite not being the underlying grammar used to produce the language, are fully compatible with that language.

Given the richness of possible lexicophonological analyses, the lack of exploration of the role of the lexicon in evaluating performance asymmetries thus raises an important question: does the asymmetry observed in the reproduceability of patterns generated by

different process interactions emerge as a consequence of how challenging it is for the underlying process interaction to be recovered given their respective surface forms, or is it because certain surface patterns are easier to reproduce *as something else* than other surface patterns? In other words, I am interested in exploring the following question in (3), again adapted from Heinz (2011).

(3) "Given a finite set of meanings paired with surface forms from a particular phonological grammar, what phonological grammar and lexicon generated them?"
Or, *do some patterns persist due to how easy is it for the learner to reproduce them in some form, either by recovering the underlying lexicon and grammar or learning a different lexicon and grammar entirely?*

For example, given the OR [ikitʃi], what is the *space* of possible lexicons and grammars that could have generated it? How does this space help or interfere with the model's ability to produce [ikitʃi] over an alternative language containing the OR [ikiti]?

Current models of phonology investigating the learnability of different process interactions have converged on the same broad conclusion: certain patterns are easier for the learner to replicate and generalize due to MISLEARNING as a result of GRAMMATICAL AMBIGUITY: some distributions are ambiguous between multiple compatible grammars that are only distinguished by a small set of forms in the data. The manner in which grammatical ambiguity affects model performance is characterized as a consequence of SEARCH; either the model takes too long to learn the correct generalization and lands on an incomplete grammar by the end of the learning period (Jarosz 2016; Prickett 2019), or the model is more likely to get caught on an alternative grammatical analysis due to local optimization (Nazarov & Pater 2017). However, the introduction of the lexicon into the learning process introduces an alternative characterization: some patterns are consistent with or nearly consistent with more grammars and lexicons than others, some of which mischaracterize the data or posit different hypotheses than the one used to generate the data originally. Under this hypothesis, the observed differences in performance emerge not from the learnability of the process interaction directly, but rather from the number of lexicons and grammars related to that pattern. I restate this in (4).

(4) HYPOTHESIS: observed asymmetries in performance emerge as a consequence of the difference in the number of joint lexicon-grammar hypotheses that are consistent or nearly consistent with certain patterns of phonological alternation over others.

In this dissertation, I aim to make the contributions in (5).

(5)
- I propose a novel NOISY-CHANNEL model of LEXICOPHONOLOGICAL LEARNING. This model jointly infers a weighted space of consistent and nearly consistent lexicons and grammars from labelled, unparsed surface data. Predictions are generated given the entirety of the inferred weighted space.
- I evaluate the model with respect to two artificial language learning experiments, which, despite involving the same underlying processes, produced contradictory results. I show that the model is able to achieve the results of both experiments under a unified account, in which the generalizability of a pattern is determined by the number of hypotheses compatible or nearly compatible with the surface distributions given to the model.
- I investigate in more detail how much the space of compatible versus nearly compatible hypotheses contribute to the model's performance. I show using two different computational experiments that the ability of the model to generalize a pattern is overwhelmingly dictated by the relative weighted number of hypotheses associated with alternative, similar distributions, with little effect of the space of hypotheses fully consistent with the data.

The remainder of the chapter will be organized in the following manner. First, I will provide the empirical background surrounding the claimed learnability asymmetry of process interactions, as well as how these asymmetries are examined computationally. I discuss how the introduction of the lexicon can affect the hypothesis space and in turn influence performance on reproducing and generalizing these patterns. I then discuss previous research on typology, arguing that all the proposals made to account for differences in typological frequency encode two basic properties: ① the number of grammars associated with a particular surface distribution, and ② the weighted similarity of that pattern to other, nearly identical patterns. I extend this idea to encompass the space of possible lexicons as well. I provide background on the noisy channel, a concept used to model variation and applied to different linguistic phenomena, including phonetic category perception and syntactic acquisition. I will use this concept in conjunction with the joint space over lexicons and grammars to quantify and encode the effect of the entire inferred weighted space of both consistent and nearly consistent lexicons and grammars on replicability. I then conclude with a roadmap of the dissertation.

1.1 Learnability of Process Interactions

It is well-known that phonological processes are capable of interacting with each other. While there are many ways of characterizing this interaction, I focus our attention on a particular class of representation: rule-based theories of phonology. Rule-based formalisms such as SPE (Chomsky & Halle 1968) represent phonological processes as rules that sequentially apply to an underlying form to produce its observed output form. Rules are traditionally defined using three parameters: ① the target, or the sound or sounds that will be changed, ② the change, or the sound that each target sound will change into, and ③ the environment, or the context the target sound must be found in in order to change. The target found within the specified environment is referred to as the STRUCTURAL DESCRIPTION. These components are demonstrated in the example given in (1), which I repeat below.

(1) a. DELETION: V → ∅ / _V
    b. PALATALIZATION: t → tʃ / _i

The grammar applies each rule sequentially, examining the input and locating all instances in which the structural description is met. For each case in which the structural description is satisfied, the target is remapped into the change. For example, the UR /ikat-a-i/ satisfies the structural description for the deletion process, as the vowel [a] is followed by another vowel [i]. The target segment is thus deleted, producing the form [ikati].

There are two dimensions in which any two processes A and B can potentially interact: ① whether A creates or eliminates the environment in which B applies, and ② whether A precedes or follows B. For example, consider the rules given in (1) above. Depending on the UR given to the rules, as well as the order in which we apply these rules, we can produce different outputs. Given an input such as /ikat-a-i/, deletion generates an environment to which palatalization can then apply. As discussed above, applying deletion to the input produces the form [ikati]. This form satisfies the structural description of the palatalization process, as the coronal consonant [t] is followed by the high vowel [i]. If deletion precedes palatalization, deletion successfully creates the environment for palatalization to then apply, generating the OR [ikatʃi]. This is known as a FEEDING interaction. In contrast, if deletion follows palatalization, deletion occurs too late in the derivation to condition palatalization, and palatalization thus fails to apply, generating the OR [ikati]. This is known as a COUNTER-FEEDING interaction.

In contrast, consider a slightly different UR /ikat-i-a/. The application of deletion in (5) now eliminates the environment for which palatalization could have applied: while the input satisfied the structural description [ti] for palatalization, deleting the vowel

[i] destroys the structural description, thus preventing palatalization from applying. If deletion precedes palatalization, deletion successfully destroys the requisite structure for palatalization to then apply, resulting in the OR [ikata]. This is known as a BLEEDING interaction. In contrast, if deletion follows palatalization, deletion occurs too late in the derivation to block palatalization, and thus palatalization is able to apply, resulting the OR [ikatʃa]. This is known as a COUNTER-BLEEDING interaction.

Note that under a rule-based system, there is no difference in representational difficulty among the four ORs; each demonstrated form can be generated under the same set of underlying atomic set of segments and rules. Any difference in performance observed under this theory, then, will not emerge from representational attributes, but directly as some property of the model. This is not necessarily true of all systems of representation, such as Optimality Theory (Smolensky & Prince 1993). Under the Optimality-theoretic formalism, certain process interactions can be distinguished by whether they can be formalized phonologically at all. Asymmetries in performance under this theory is complicated by properties of the representation rather than properties of the model. I return to the consequences of the selected phonological theory in Chapter 5.

1.1.1 Theoretical and empirical background

It has been claimed that some process interactions are easier to learn than others. The basis for this assertion has been observations in two empirical domains: the historical domain and the experimental domain. In the historical domain, it has been argued that some process interactions are more prone to changing across generations of speakers than others, exhibiting surface phenomena in which processes seemingly are lost or reordered (Hansson & Sprouse 1999; Kiparsky 1965, 1968, 1971; O'Bryan 1974). In the experimental domain, it has been demonstrated that adult participants have a more difficult time extending generalizations observed in toy languages generated from certain process interactions compared to others (Ettlinger 2008; Kim 2012; Brooks, Pajak, & Baković 2013; Prickett 2019).

In the historical domain, Kiparsky demonstrated that patterns generated from certain process interactions were more likely to change in later generations into patterns exhibiting the opposite order. He proposed in sequence two different learnability biases in order to characterize these tendencies: the ① MAXIMUM UTILIZATION and ② TRANSPARENCY biases.

The Maximal Utilization bias places the direction of asymmetry primarily on the notion of whether the interaction utilized both processes or not. This is outlined in (6).

(6) MAXIMAL UTILIZATION (Kiparsky 1968)
Prefer orderings in which rules are most maximally applied.

Under this bias, the feeding and counter-bleeding interactions would be expected to be easier for the learner to acquire than the counter-feeding or bleeding interactions.

The Transparency bias emphasizes that it is the surface informativity of the forms generated from the different process interactions that serves as the main contributor to the learning asymmetries. The definition of the Transparency bias is defined in (7).

(7) TRANSPARENCY (Kiparsky 1971; McCarthy 1999)
For a rule of structure A → B / C_D to be transparent, it must not exhibit:

a. UNDERAPPLICATION: the structure A is found in the environment C_D on the surface, but the rule does not apply, i.e. = NOT SURFACE-TRUE.

b. OVERAPPLICATION: the structure B is created by the rule, but the environment C_D is no longer visible on the surface, i.e. = NOT SURFACE-APPARENT.

Rules that exhibit one of the above are OPAQUE; prefer orderings that are transparent.

Defined originally in terms of a single rule within an interaction, the properties outlined in (7) have since been used to describe process interactions as a whole. For example, the input /ikat-a-i/ and rule order ⟨PAL., DEL.⟩ form a counter-feeding interaction that meets the criterion for being not surface-true: the interaction when applied to this input produces an output in which the environment for palatalization is met on the surface, but the process does not apply, due to the conditioning environment resulting from deletion occurring too late. Thus, it appears on the surface as though palatalization underapplies. The input /ikat-i-a/ and identical rule order ⟨PAL., DEL.⟩ in contrast form a counter-bleeding interaction that meets the criterion for being not surface-apparent: the interaction produces an output in which palatalization has seemingly applied, but the environment is no longer met on the surface, due to deletion having eliminated the environment. Thus, it appears on the surface as though palatalization overapplies. In contrast, the feeding and bleeding interactions are transparent: it is clear on the surface why a process applied, or failed to apply. Thus, under this bias, the feeding and bleeding interactions are predicted to be easier for the learner to acquire than the counter-feeding or counter-bleeding interactions.

Interestingly, the two biases overlap in terms of which interactions are posited to be easier or harder to learn: both the Maximal Utilization and Transparency biases favor the feeding interaction over the counter-feeding interaction, while the preference for the bleeding or counter-bleeding interactions differ depending on the bias. Thus, in historical terms, the directionality of change under both biases is not necessarily predictable from the biases dictated in (6) and (7) above. This is visually demonstrated in Table 1-1.

The empirical backdrop supporting these biases in the historical context, however, are

Table 1-1: Joint effect of the Maximum Utilization and Transparency biases. Darker shaded cells correspond to languages predicted to be relatively more difficult to learn.

		Maximal Utilization	
		√	×
Transparency	√	Feeding	Bleeding
	×	Counter-bleeding	Counter-feeding

far from consistent; for as much evidence there has been for certain interactions to be more likely to change historically than others (Kiparsky 1965, 1968; King 1969), there have also been many cases in which those interactions are shown to persist or even be preferred to their opposite ordering (Kaye 1974; Kenstowicz & Kisseberth 1971; Kiparsky 1971; King 1973b).

In the experimental domain, while there is indeed an observed empirical difference in performance in generalizing different underlying process interactions, the preference likewise did not necessarily follow any consistent direction either: Ettlinger (2008) found a preference in performance for the counter-feeding interaction over the counter-bleeding interaction; Kim (2012) found a preference for the counter-feeding interaction over the feeding interaction; Brooks and colleagues (2013) found a preference in favor of non-interaction or non-application; and Prickett (2019) observed that participant performance depended not only on what underlying interaction was being learned, but also which form was being evaluated, observing evidence of both the Maximal Utilization and Transparency bias, depending on the form being tested. Of particular note is that while the Kim and Prickett experiments examined the learnability of process interactions involving the same underlying processes, the observed outcomes were in complete opposition to one another.

It is thus clear that the direction of asymmetry is *not* consistent. Curiously, both the historical and experimental results indicate that while the *directionality* is not seemingly predictable based on the biases given above alone, there is indeed an observable difference in performance crosscutting different process interactions. I summarize in (8) the broad empirical observations provided above into three basic points.

(8) a. Patterns generated from some process interactions are more likely to be changed diachronically or harder to reproduce and generalize than others.
b. The directionality of the asymmetry is not consistent.
c. All phonological interactions can be productive and can be innovated.

Other competing biases have been proposed in order to account for the variability in

results. For example, it has been independently proposed that language is biased towards distributions of increasing uniformity. This has been commonly referred to as Paradigm Uniformity (**albright2005morphological**; King 1973a; Steriade 2000). Under this bias, regardless of the underlying process interaction used to generate the pattern, a surface pattern is more likely to transition to a state of greater uniformity than one of greater contrast (cf. Kisseberth 1973 and the effect of semantic contrast in maintaining surface contrasts.). The introduction of these three biases, then, has been argued to be the driving force behind the empirical observations.

While these formal characterizations have provided a good starting point for predicting *why* certain asymmetries are more likely to appear (King 1973a), they do not provide a reasoning as to *how* these definitional properties result in the observed empirical asymmetries. I explore some recent computational work that has been done precisely to tackle this question in the following section.

1.2 The Role of the Grammar in Learning

The computational literature investigating the presence and manifestation of learnability biases in general has been interested in determining whether empirical asymmetries emerge as a *consequence* of the learning process (i.e. Learnability or Channel Bias, Staubs 2014; Bane & Riggle 2008; Stanton 2016, Ohala 1992; Blevins 2004) or as an innate, hard-coded bias (i.e. Substantive Bias, Wilson 2006; Moreton & Pater 2012; White 2017). Most work done on the learnability of process interactions have primarily focused on the effect of the learning process on achieving the observed results. Most of these computational models has converged on the same underlying conclusion: differences in performance emerge as a result of the differences in the distribution of surface data, and the ability of the model to recover the underlying grammar given these distributions, where some surface patterns provide clearer evidence for the existence of certain interactions than others (Jarosz 2016; Nazarov & Pater 2017; Prickett 2019; Prickett & Jarosz 2021; Yang & Ellis 2021). Since the space of computational models is vast and quite variable, I will present these models more schematically and impressionistically. Details for the actual implementations of each model can be found in their respective papers.

Most computational research assesses grammatical learnability in the following manner: given a set of UR-OR pairs, infer a grammar and evaluate how well that grammar predicts the expected output forms for inputs not given in training. For example, consider the sample dataset in Table 1-2, adapted from Baković (2011), which I henceforth refer to as the revised Baković languages. The data is distributed into two partitions: ① the training

Table 1-2: Sample feeding and counter-feeding dataset based on Baković (2011) illustrating the typical learning problem. Within a sub-table for each language, the left and right columns correspond to the input and output form, respectively. The outlined cells correspond to the UR-OR pairs distinguishing the two languages.

		Feeding Language				Counter-feeding Language			
		Alternating		Non-alternating		Alternating		Non-alternating	
Training		/ikit/	[ikit]	/ikik/	[ikik]	/ikit/	[ikit]	/ikik/	[ikik]
		/ikita/	[ikita]	/ikika/	[ikika]	/ikita/	[ikita]	/ikika/	[ikika]
		/ikiti/	[ikiti]	/ikiki/	[ikiki]	/ikiti/	[ikiti]	/ikiki/	[ikiki]
		/ikitai/	[ikitʃi]	/ikikai/	[ikiki]	/ikitai/	[ikiti]	/ikikai/	[ikiki]
Testing		/akit/	???	/akik/	???	/akit/	???	/akik/	???
		/akita/	???	/akika/	???	/akita/	???	/akika/	???
		/akiti/	???	/akiki/	???	/akiti/	???	/akiki/	???
		/akitai/	???	/akikai/	???	/akitai/	???	/akikai/	???

data and ② the testing data. The training data, as the name suggests, are the data used to train the model, and provides information of both the underlying form and its respective observed OR. For example, the input for the training datum /ikiti/ is mapped onto the output [ikitʃi]. Examining the data in the both the feeding and counter-feeding language, we observe two phenomena in action: ① the deletion of a vowel when found before another vowel underlyingly, i.e. /ikikai/ $\xrightarrow{\text{DEL}}$ [ikiki], and ② the palatalization of a coronal consonant when found before a high vowel, i.e. /ikiti/ $\xrightarrow{\text{PAL}}$ [ikitʃi]. We also observe evidence of their interaction: for the feeding language, the UR-OR pair ⟨/ikitai/, [ikitʃi]⟩ illustrates that deletion must precede palatalization /ikitai/ $\xrightarrow{\text{Del.}}$ [ikati] $\xrightarrow{\text{Pal.}}$ [ikitʃi], whereas for the counter-feeding language, the UR-OR pair ⟨/ikitai/, [ikiti]⟩ illustrates that palatalization must precede deletion /ikitai/ $\xrightarrow{\text{Pal.}}$ [ikatai] $\xrightarrow{\text{Del.}}$ [ikiti]. Given the space of possible grammars and the space of observed input-output pairs, the model is tasked with trying to determine the best grammar that maps each of the inputs to their respective outputs.

Once a grammar has been inferred given the training data, this grammar is then evaluated by providing the model with a set of testing data, indicated by the shaded cells and undefined output forms "???" in the table above. If the model successfully learns both the deletion and palatalization processes in order, the predicted outputs of the testing data would generate an output consistent with the underlying grammar, e.g. for the feeding language, /akitai/ $\xrightarrow{\text{Del.}}$ [akati] $\xrightarrow{\text{Pal.}}$ [akitʃi]. If the model does not successfully learn one of the processes or the interaction, the predicted output given the input is not borne about.

Many models of phonological learning assume gradual inference over a probabilistic

grammar, in which input-output pairs are assigned probabilities based on how close the predicted form under the grammar matches the actual form observed in the data. This is known as the LIKELIHOOD. Inference is typically performed in these models by examining the data one at a time and adjusting the parameters over the mappings incrementally based on which parameterization maximizes the likelihood of the training data. However, different parameterizations can produce similar likelihoods for the same form. For example, the counter-feeding form ⟨/ikitai/, [ikiti]⟩ can be generated from a mapping in which both the palatalization and deletion processes are learned, as above, or from a mapping in which only deletion is learned: /ikitai/ $\xrightarrow{\text{DEL}}$ [ikiti]. This is in contrast to the feeding interaction, in which the output can only be generated through the correct sequence of deletion and palatalization applying. In other words, certain forms provide ambiguous and contradicting information about the existence of a process or interaction.

The models of grammatical learnability presented above perform well in capturing certain empirical asymmetries: Jarosz (2016) was able to detect evidence of both the Maximal Utilization and Transparency biases under her Expectation-Driven Learning model (Jarosz 2015) depending on the relative frequencies of different forms given to the model; Prickett (2019) utilized the same model and found that these biases emerge even under a uniform distribution over forms, so long as one examines the individual types of forms being generated; Nazarov and Pater (2017) found with the Stratal Maximum Entropy learner that transparent interactions were more likely to successfully converge on the correct grammar than their opaque counterparts; and Rasin (p.c.) discovered that his Minimum Description Length model (Rasin et al. 2015) took longer to converge on a counter-bleeding interaction compared to its bleeding complement; and Yang and Ellis (2021) found not only a Maximum Utilization bias under a similar Minimum Description Length model (Ellis et al. 2022), but also evidence suggesting that the individual processes involved in an interaction may also affect generalizability. However, these models do not address one or more of the following:

(9)
- Many of the models do not examine the effect of the lexicon on performance.
- Many of the models have not investigated how introducing these lexicons and grammars would influence performance on patterns generated from different process interactions.
- None of the models have attempted to explore the full spectrum of observed empirical asymmetries; for example, the contradictory results observed in Kim (2012) versus Prickett (2019).

In the following sections, I will discuss how each of the former two points can affect the results, and how to accomodate for them in a computational model.

Table 1-3: Sample no-voicing language from Jarosz (2016).

tak	$\langle \mu_1 \rangle$	taka	$\langle \mu_1, \mu_A \rangle$

1.3 The Role of the Lexicon in Learning

In this section, I will present previous research examining the effect of the lexicon on shaping model performance. I will then discuss specifically how the lexicon can influence the ability to replicate and generalize patterns generated from different process interactions. As before, I explore the insights of previous research using schematizations of each model. I refer the reader to each paper for specifications of each model.

There is evidence to suggest that lexical learning expands the space of compatible hypotheses given the same set of data. For example, Jarosz (2006) proposed a probabilistic Optimality-theoretic model of learning that jointly learns the underlying forms and grammar given a set of unparsed surface form and meaning pairs. This model, like many others, adopts the standard assumption by which a morpheme corresponds to a single underlying form across all contexts. Alternations observed in the data must thus emerge as a consequence of the grammar. The goal of the model is to infer a hypothesis that assigns a probability distribution over the underlying forms and grammars given the data.

Consider the following schematic dataset in Table 1-3. This language corresponds to a simple language in which no voiced obstruents are allowed to surface, and is compatible with multiple different hypotheses. For example, this output distribution can be easily accounted for using the process in (10). This process simply prevents all underlying voiced consonants from surfacing as voiced.

(10) DEVOICING: D → T

This phonological process, however, is compatible with multiple different underlying form hypotheses. I list two such hypotheses in (11) below.

(11) Two possible hypotheses compatible with the no-voicing language.

UNDERLYING FORM	/tat/	/tat-e/	/tad/	/tad-e/
D → T	–	–	[tat]	[tate]
OBSERVED FORM	[tat]	[tate]	[tat]	[tate]

As can observed, the same grammar is compatible with multiple underlying forms, and can result in the same output. When training the model on this small dataset, Jarosz

Table 1-4: Sample counter-bleeding language from Yang and Ellis (2021).

tat	$\langle\mu_1\rangle$	tate	$\langle\mu_1,\mu_A\rangle$	tatʃi	$\langle\mu_1,\mu_B\rangle$	tatʃee	$\langle\mu_1,\mu_B,\mu_A\rangle$
kat	$\langle\mu_2\rangle$	kate	$\langle\mu_2,\mu_A\rangle$	katʃi	$\langle\mu_2,\mu_B\rangle$	katʃee	$\langle\mu_2,\mu_B,\mu_A\rangle$
tatʃ	$\langle\mu_3\rangle$	tatʃe	$\langle\mu_3,\mu_A\rangle$	tatʃi	$\langle\mu_3,\mu_B\rangle$	tatʃee	$\langle\mu_3,\mu_B,\mu_A\rangle$
kak	$\langle\mu_4\rangle$	kake	$\langle\mu_4,\mu_A\rangle$	kaki	$\langle\mu_4,\mu_B\rangle$	kakee	$\langle\mu_4,\mu_B,\mu_A\rangle$

indeed found this very effect: the model, while learning the correct grammar, had uniform probability over possible underlying forms for the stem: $P(/\text{tat}/) = P(/\text{tad}/) = P(/\text{dat}/) = P(/\text{dad}/)$. This observation has been mirrored in other theoretical and computational formalisms as well (Nelson 2019; Pater et al. 2012). In other words, incorporating lexical learning expands the space of possible hypotheses compatible with the data by allowing for multiple underlying form hypotheses for the same grammar.

The lexicon also has been demonstrated to allow for alternative grammatical analyses compatible with the data. Yang and Ellis (2021) adopted a Minimum Description Length model in order to explore how performance on different process interactions is influenced not only by the lexicon and grammar, but also by the formal properties of the individual processes used to express the language. This model allows the lexicon to consider exceptions by memorizing allomorphs seen on the surface. Given a set of surface form and meaning pairs, the model is tasked with determining the optimal lexicon and grammar that minimizes the combined cost of encoding each.

Consider the toy counter-bleeding language in Table 1-4.[2] The language was generated from two underlying processes, outlined in (12). The first process corresponds to the familiar palatalization process, in which [t] palatalizes to [tʃ] when preceding [i]. For example, the form [tatʃi] was generated from the input /tat-i/: /tati/ $\xrightarrow{\text{Pal.}}$ [tatʃi]. The second process corresponds to a vowel harmony process, in which a vowel assimilates in height to the following vowel. For example, the form [kakee] was generated from the input /kak-i-e/: /kak-i-e/ $\xrightarrow{\text{VH}}$ [kakee]. The counter-bleeding interaction is formed by applying palatalization first, followed by vowel harmony. Thus, the form [tatʃee] is produced from the underlying form /tat-i-e/: /tat-i-e/ $\xrightarrow{\text{Pal.}}$ [tatʃie] $\xrightarrow{\text{VH}}$ [tatʃee].

(12) a. Vowel Harmony: V → [αhigh] / _[αhigh]
b. Palatalization: t → tʃ / _i

While the original dataset is generated via the aforementioned lexicon and interaction,

[2] The toy language here is not identical to the one presented by Yang and Ellis, who instead utilized a voicing process rather than a palatalization process. This change was made purely for presentational reasons, and will not alter the overall conclusion.

it is not the only hypothesis consistent with the data. For example, consider the alternative hypothesis in (13). Here, instead of a palatalization process, a deletion process is inferred instead, wherein a consonant deletes before another consonant. Thus the form [tatʃee] is generated not through a counter-bleeding interaction, but via a non-interaction, with the [t]~[tʃ] alternation in the original analysis becoming a [tʃ]~[∅] alternation instead. Consequently, the posited URs differ from the original formulation as well, with the /-i/ suffix now incorporating the palatalized consonant in its input /-tʃi/. While an extreme case, this hypothesis is fully consistent with the data given; forms that are not compatible with the lexicon and grammar, such as the non-alternating [kaki] and [kakee], can simply be encoded such that the alternation is encoded in the lexicon rather than the phonology.

(13) Alternative lexicon and grammar for the counter-bleeding language in Table 1-4

Underlying Form	/tat/	/tat-e/	/tat-tʃi/	/tat-tʃi-e/
C → ∅ / _C	–	–	[tatʃi]	[tatʃie]
V → [αhigh] / _[αhigh]	–	–	–	[tatʃee]
Observed Form	[tat]	[tate]	[tatʃi]	[tatʃee]

This is exactly what was observed in their model; while the *most optimal* – i.e. hypothesis that globally minimizes the encoding cost of both the lexicon and grammar – was indeed found to be the original underlying hypothesis used to generate the language, the model found that hypotheses of the sort as in (13) were optimal under a slightly different parameterization of the model favoring shorter rule encodings.

Important to note is that different hypothesis can potentially make different predictions on held-out data. For example, suppose that in addition to the data given in Table 1-4, the model is given the OR [kek]. In order to minimize the size of the lexicon, the model would adopt the UR hypothesis /kek/. If we take this UR and concatenate the suffix /-tʃi/, as discussed above, the model would predict the output [ketʃi], which does not correspond to the predicted output under the intended lexicon and grammar, i.e. [keki]. Thus, while incorporating a richer lexicon may augment the space of possible grammars that are compatible with the given data, the predictions for each hypothesis on held-out data may vary vastly.

In the discussion above, I highlighted two important ways in which inference over both the lexicon and grammar can influence the generalizability of a pattern: ① the lexicon expands the space of compatible hypotheses, which in turn allows for completely new analyses for the same data, and ② different analyses under the space can result in completely

different predictions on held-out data compared to the intended underlying grammar. In what follows, I discuss how these factors can be applied to predict the generalizability of different process interactions.

I demonstrated that certain patterns are compatible with more lexicon-grammar pairs than others. Recent work by Baković and Blumenfeld (2018; 2022) sought to formally characterize process interactions based on the relationship between each process and the forms they produce for the other. They proposed two atomic properties: ① INPUT REMOVAL AND PROVISION and ② OUTPUT REMOVAL AND PROVISION, defined in (14) and (15) below.

(14) INPUT INTERACTIONS: given two phonological processes A and B:

a. A input-provides B if there exists from mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $x \xrightarrow{B}$ is vacuous
   (iii) $y \xrightarrow{B}$ is not vacuous
   A creates a form in which B can apply where before, it could not.

b. A input-removes B if there exists some mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $x \xrightarrow{B}$ is not vacuous
   (iii) $y \xrightarrow{B}$ is vacuous
   A creates a form in which B cannot apply where before, it could.

(15) OUTPUT INTERACTIONS: given two phonological processes A and B:

a. A output-provides B if there exists some mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $\xrightarrow{B} x$ is vacuous
   (iii) $\xrightarrow{B} y$ is not vacuous
   There is a form y that A can create that B can also create without applying A.

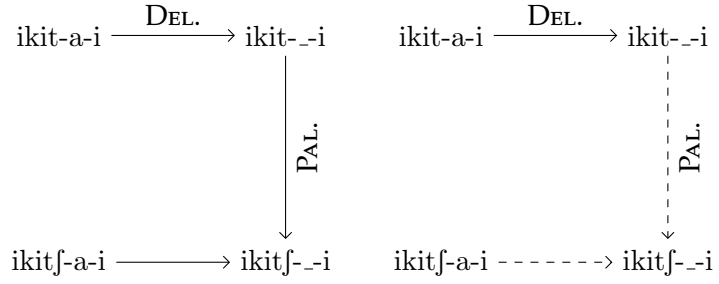
b. A output-removes B if there exists some mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $\xrightarrow{B} x$ is not vacuous
   (iii) $\xrightarrow{B} y$ is vacuous
   There is a form y that B can create that A cannot create without first applying B.

While these properties are formalized over the mappings, these definitions are directly related to the space of compatible lexicons and grammars.In order to demonstrate the relationship between these properties and their respective impact on the lexicophonological

space, consider the process interactions for the deletion and palatalization processes presented in (1), and the respective feeding and counter-feeding ORs [ikitʃi] and [ikiti] under the revised Baković languages presented in Table 1-2. I graphically map out the space of possible compatible inputs and grammars for each OR in (16). URs and mappings are organized into vertices and edges, respectively. The edges from left to right correspond to the nonvacuous application of the deletion process, while the edges from top to bottom correspond to the nonvacuous application of the palatalization process. Each vertex corresponds to a possible form that serves as an input to a non-vacuous mapping or output of a nonvacuous mapping. For example, the form [ikiti] can either serve as a nonvacuous output of deletion, i.e. [ikitai] $\xrightarrow{\text{Del.}}$ [ikiti], or nonvacuous input to palatalization, i.e. [ikiti] $\xrightarrow{\text{Pal.}}$ [ikitʃi].

The edges connecting two vertices correspond to the minimal derivational path to get from one form to another. For example, given the pair of forms ⟨/ikitai/, [ikitʃi]⟩, the minimum set of derivations needed to get from the input to its correspondent output is a deletion process followed by a palatalization process, or ⟨Del., Pal.⟩. Solid edges correspond to paths that eventually lead to the OR for the given language, while dashed lines correspond to paths that lead to an alternative output. For example, the OR belonging to the feeding language [ikitʃi] can be generated a number of different ways given the deletion and palatalization processes, such as via: ① the UR /ikitai/ provided deletion applies followed by palatalization, i.e. /ikitai/ $\xrightarrow{\text{Del.}}$ [ikiti] $\xrightarrow{\text{Pal.}}$ [ikitʃi], ② the UR /ikiti/ provided palatalization applies, i.e. [ikiti] $\xrightarrow{\text{Pal.}}$ [ikitʃi], and ③ the UR /ikitʃai/ provided deletion applies, i.e. /ikitʃai/ $\xrightarrow{\text{Del.}}$ [ikitʃi]. In contrast, the OR [ikiti] belonging to the counter-feeding language, can only be generated a few ways, such as via the UR /ikitai/ provided deletion applies and palatalization does *not* follow, i.e. /ikitai/ $\xrightarrow{\text{Del.}}$ [ikiti]. Applying palatalization after deletion produces the unintended OR *[ikitʃi], instead. This visualization thus lays out the minimum set of requirements on the grammar to produce a particular OR given each UR.

(16) Compatible lexicons and grammars for the feeding form [ikitʃi] (LEFT) and counter-feeding form [ikiti] (RIGHT)



The graph above visualizes the effect the input and output interactions have on the space of compatible lexicon-grammar pairs. Crucially, it is clear that the ORs for each respective process interaction are compatible with a different number of lexicons and grammars. In order to illustrate how the space of interacting lexicons and grammars can in principle influence the behavior of the model given different outputs, I will provide only an impressionistic view of the effect here, focusing our attention on the contrast observed between the feeding and counter-feeding interactions. I will revisit and expand on this discussion to include the bleeding and counter-bleeding interactions in Chapter 3.

I proposed that a pattern is more likely to persist if there are more distinct lexicons and grammars that are compatible with it. We can visually confirm that there are, as predicted, more ways of combining the lexicon and grammar to generate the feeding OR [ikitʃi] than the counter-feeding OR [ikiti]. This emerges directly as a consequence of the properties explicated above. In (17) and (18), I provide sample form pairs that demonstrate the relevant properties of each input and output interaction for the feeding and counter-feeding ORs.

(17) a. Deletion input-provides Palatalization: $\langle x, y \rangle$ = ⟨ikitai, ikiti⟩
   (i) $x \xrightarrow{\text{Pal.}}$ is vacuous: /ikitai/ $\xrightarrow{\text{Pal.}}$ [ikitai]
   (ii) $x \xrightarrow{\text{Del.}}$ is not vacuous: /ikitai/ $\xrightarrow{\text{Del.}}$ [ikiti]
   (iii) $y \xrightarrow{\text{Pal.}}$ is not vacuous: /ikiti/ $\xrightarrow{\text{Pal.}}$ [ikitʃi]

(18) a. Deletion output-provides Palatalization: $\langle x, y \rangle$ = ⟨ikitʃai, ikitʃi⟩
   (i) $x \xrightarrow{\text{Del.}}$ is not vacuous: /ikitʃai/ $\xrightarrow{\text{Pal.}}$ [ikitʃi]
   (ii) $\xrightarrow{\text{Pal.}} x$ is vacuous: /ikitʃai/ $\xrightarrow{\text{Pal.}}$ [ikitʃai]
   (iii) $\xrightarrow{\text{Pal.}} y$ is not vacuous: /ikiti/ $\xrightarrow{\text{Pal.}}$ [ikitʃi]

Because deletion input-provides and output-provides palatalization, different ORs are compatible with different lexicon-grammar pairs. Specifically, certain hypotheses are available for the outputs generated from the ordering ⟨Del., Pal.⟩, i.e. the feeding interaction, that are not available or blocked for the outputs generated from the opposite ordering ⟨Pal., Del.⟩, i.e. the counter-feeding interaction. Because the feeding interaction is input-providing, the output of deletion can be used as an alternate UR for the feeding form without losing the ability to posit both processes: /ikiti/ $\xrightarrow{\text{Del.}}$ [ikiti] $\xrightarrow{\text{Pal.}}$ [ikitʃi] Similarly, because the feeding interaction is output-providing, alternative URs such as /imatʃai/ in which the [t]~[tʃ] alternation is encoded into the lexicon is also available to be used to generate the output, again without losing the availability to posit both processes: /ikitʃ-a-i/ $\xrightarrow{\text{Del.}}$ [ikitʃi] $\xrightarrow{\text{Pal.}}$ [ikitʃi]. For the counter-feeding form, these options are not possible; the UR /ikiti/ cannot be used to generate the output unless palatalization is lost: /ikiti/ $\xrightarrow{\text{Del.}}$ [ikiti] $\xrightarrow{\text{Pal.}}$ *[ikitʃi]. Likewise,

there are no alternate URs that can be posited in which both deletion and palatalization are encoded lexically without also losing the ability to posit the palatalization process.

I have demonstrated in this section that the introduction of the lexicon results in an expansion of the hypothesis space; not only does incorporating the lexicon expand the hypothesis space over compatible grammars, each of which may generate wildly different predictions, this inferred space may also differ depending on the distribution of the data given to the model. Thus, the observed asymmetry in performance observed between different process interactions may fall out as a by-product of the differences in the space of hypotheses given the different distributions generated by each process interaction as opposed to incomplete learning.

1.4 Typology, Similarity, and Grammatical Spaces

Research investigating the relationship between the phonological grammar and typology has primarily been interested in one of two potentially overlapping topics: ① determining why some phonological patterns are attested while others are not, and ② determining why some attested phonological phenomena are more frequent than others. The basic principle underlying all proposals related to this topic is as follows: patterns are more likely to be typologically frequent if ① they are more likely to be *innovated* than other patterns, and ② if the pattern is more likely to *persist* over time than alternative patterns. I discuss a few of these proposals below.

Bane and Riggle (2008) examined what factors are correlated with whether a given stress pattern is attested versus unattested, as well as the typological frequency of each attested pattern. They noted three possible factors, two of which I discuss here. The first property corresponds to the notion of the R-VOLUME (Riggle 2010), or the relative number of unique grammars compatible with a given pattern. Typological frequency could in principle emerge as a consequence of the redundency of a pattern: given indeterminate data, the form consistent with more grammars occupies more of the probability space over possible output forms than alternative forms and thus is more likely to be chosen. The second property corresponds to the notion of CONFUSABILITY, or how much of what kind of data is necessary in order to disambiguate a pattern from other, related distributions. Here, typological attestedness emerges as a consequence of the distinctness of a pattern; attested patterns require less disambiguating data in order to identify that unattested patterns, and thus are more likely to persist than other patterns. Both properties highlight two distinct properties outlining typology: the prior probability a particular pattern possesses given a space of grammars, and how similar that pattern is to other related patterns.

The same properties explicated above are echoed in research by Rafferty et al. (2013), who found that the persistence of a phenomenon, and thereby its typological frequency, are modulated by the interaction of three factors: ① the learnability of that phenomenon, or how likely a pattern exhibiting that phenomenon is to be learned, ② the number of patterns consistent with that phenomenon, and ③ how similar a pattern exhibiting the phenomenon is to other alternative distributions[3]. They demonstrate this experimentally by training participants on a vowel harmony pattern, a typologically common phonological occurrence amongst the world's languages (Hulst 2016). They performed two experiments. In the first experiment, participants were taught one of two languages: a front harmony phenomenon, in which a vowel agrees in backness with the previous vowel [pelit] and [bisit], as well as a front-height harmony phenomenon, where a vowel is front if preceded by a mid vowel and back if preceded by a high vowel [pelit] and [bisut]. They were then evaluated on how well they generalized the pattern seen in the language to unfamiliar forms. In the second experiment, participants were tasked with transmitting the artificial language over multiple generations. A set of participants, representing a generation of speakers, were trained on the vowel harmony language and, as in the previous experiment, tested on how well they generalized the pattern obsered in training to unfamiliar testing forms. The forms produced by the participants were then used as the training data for the next set of participants. They discovered that while participants in the first experiment performed better at generalizing the attested vowel harmony pattern over the unattested front-height harmony pattern, participants in the transmission experiment quickly lost the vowel harmony within two generations of speakers; thus, although vowel harmony is more learnable, it does not necessarily mean that the pattern will persist over generations.

They demonstrated via a simple transmission model that this result emerged due to two factors. First, they noted that the number of surface patterns that are consistent with a vowel harmony phenomenon is much smaller than the surface patterns consistent with non-harmony phenomenon. Thus, even through a phenomenon is more learnable, if it occupies only a small portion of the hypothesis space, if the phenomenon is ever lost, the chances of innovating it again are extremely small. Second, they argued that if a pattern is too similar in form to another, then the pattern may fail to persist over generations of speakers as it is learned as something nearly consistent, but crucially *distinct* from the original pattern. Thus, even through a phenomenon is more learnable, if it is similar to an

[3] Note that Rafferty and colleagues define a phenomenon as a distribution over patterns. In other words, they classify patterns in terms of whether they exhibit what a particular linguistic phenomena, such as front harmony, e.g. CiCi versus a different phenomenon such as disharmony, e.g. CiCu *or* CuCi. The conclusions drawn from their experiments are still applicable to us, with phenomenon instead referring to the set of all languages that produce the same predictions on observed data, and a pattern an element of that set.

alternative phenomenon, it may be prone to being mislearned.

In addition to the hypothesis space and similarity of a pattern to another, there has been independent work examining the relationship between learnability and both typological attestedness and frequency. Patterns that appear more frequently in language, if at all, emerge as a consequence of those patterns being easier to be learn, whether it be from innate biases (Moreton 2008) or as a consequence of the learning process (Stanton 2016; Staubs 2014; White 2017). Stanton (2016) sought to relate learnability to typological attestedness in various stress patterns. In particular, Stanton explored whether midpoint pathologies, a class of unattested stress patterns, are unattested not because the theory prevents these distributions from being represented, but instead as a consequence of learnability principles. Using the Gradual Learning Algorithm (Boersma & Hayes 2001; Magri 2012), Stanton evaluated how long it took for the model to converge on various attested stress patterns, such as languages in which stress is assigned to the antepenultimate syllable, versus unattested patterns, such as various midpoint pathologies. One key finding was that, given a distribution of words with relatively short length, i.e. less than five syllables, certain midpoint patterns could never be learned. This was due to the fact that midpoint patterns often required a superset of the conditions neceesary to produce the attested patterns, and ambiguous data do not provide enough information to acquire the necessary conditions to generate the unattested pattern. The basic story behind the learnability accounts thus remains the same: certain patterns require fewer necessary grammatical conditions to learn, and, in the absence of disambiguating data, certain patterns are favored over others.

The discussion above indicates that there is some relationship between the space of possible hypotheses, learnability, and the typology. Initial discourse seems to suggest that the hypothesis space poses a formidable influence on the persistence of a pattern over time. In order to capture the tendencies observed with respect to process interactions, then, a model that jointly encodes the *learnability* of a pattern, the *space* of hypotheses compatible with a pattern, and the *similarity* of that pattern to other patterns is ideal. The space of consistent hypotheses is readily accounted for by way of the multiple lexicon and grammar approach of analysis; how is learnability and similarity encoded? In the next section, I present one way of quantifying similarity by way of the noisy channel.

1.5 Noisy Channel Models and Learnability

I make the claim that learners, when given a set of data, carry a level of uncertainty in the reliability of the data and as a consequence the hypotheses used to generate the data. The idea that learners' behavior reflects the combined influence of many of these different

hypotheses is consistent with a broad class of Bayesian models known as NOISY CHANNEL MODELS (Feldman & Griffiths 2007; Levy 2008). These models assume that the data learners observe may have been corrupted by a noise process, which creates additional uncertainty about what the uncorrupted data look like. Distrust in data allows the model to consider hypotheses that nearly fit the data, but also allows it to ignore certain forms if they do not show up as often. These models have been applied to a variety of linguisic phenomena, including phonetic category learning (Feldman & Griffiths 2007), syntactic parsing (Levy 2008), and syntactic acquisition (Perkins, Feldman, & Lidz 2017; Schneider, Perkins, & Feldman 2020).

The noisy channel defines a method of calculating the likelihood of the set of ORs given the predictions of the model. Given a lexicon and grammar hypothesis, the model generates a predicted output, labelled the expected form, or ER. The noisy channel defines a distribution that computes how similar the set of ERs corresponds to the actual ORs observed in the data: the more similar the ERs are to the ORs, the higher the likelihood. The noisy channel is often incorporated into a larger Bayesian model, where the output of inference is a POSTERIOR over the entire inference space. The posterior distribution corresponds to a weighted distributions over likely hypotheses given the data, combining all three main components noted by Bane and Riggle, as well as Rafferty and colleagues.

The posterior assigns probabilities over all possible hypotheses given the set of data. Hypotheses that capture the data better tend to have higher posterior probabilities than those that do not. As the likelihood contributes to the overall posterior probability of a given hypothesis, hypotheses that generate similar predictions produce similar likelihood, and generally possess the same posterior probability. However, the posterior probabilities assigned to each individual hypothesis are modulated by the number of patterns each collective set of hypotheses predict. This reflects the intuition underlying the notion of the grammatical space, as well as similarity: the more hypotheses that generate a particular pattern, the great the *a priori* weight of that pattern, and the more similar a pattern is to the observed pattern, the greater respective posterior assigned to each associated hypothesis is. In other words, even though a single lexicon and grammar is assigned very high posterior probability, if the language it produces is associated with a small hypothesis space, it can be overshadowed by alternative languages associated with more hypotheses.

In addition to the likelihood, the posterior probability is influenced by the PRIOR, or the pre-existing beliefs about which hypotheses are innately more probable. As discussed above, while the use of context-specific URs has been noted and used to analyze various linguistic phenomenon in the literature, most computational models assume a single UR for a given meaning across all contexts it is found in. If we adopt the same assumption,

the breadth of hypotheses associated with a particular output pattern is thus tempered by how many unique URs a lexicon posits.

This posterior space is then used to generate predictions on unfamiliar, held-out forms. Under this hypothesis, patterns are more likely to be able to be produced as a consequence of not only the number of lexicons and grammars compatible with the data, but also the number of lexicons and grammars that *nearly* capture the data. Depending on the distribution of forms, which are influenced by the underlying process interactions used to generate them, we observe differences in performance.

1.6 Outlining the Rest of the Dissertation

The rest of the dissertation will be organized as follows. I first go over the formalization of our noisy channel learner, combining the elements discussed in this chapter into a working model. I next evaluate this model with respect to two experimental results reported in the literature, and demonstrate this model is able to generate the observed empirical asymmetries. I conclude with a deep dive into the components of the model, as well as areas of improvement.

Chapter 2

Defining the Lexicophonological Space: the Noisy-Channel Model

In this chapter, I walk through the formalization of the noisy-channel lexicophonological learner. The chapter is distributed into four sections. ① I discuss in detail what knowledge the model has access to prior to learning, such as the inventory of segments and features, as well as the space of observed surface forms and possible meanings. ② I go over how the model builds the lexicon and generates the observed surface forms. ③ I explain how the model performs joint inference over the lexicon and grammar using Markov Chain Monte Carlo estimation. ④ I discuss how this inferredspace encodes the notion of the lexicophonological space and how it in turn is used to generate predictions. I focus our discussion on rule-based grammars, e.g. SPE (Chomsky & Halle 1968), but the model can be applied to any phonological theory that maps underlying forms to surface forms. I reassess this assumption in Chapter 5.

2.1 Assumptions and Data Structure

I assume that the model has access to two pieces of information by the time morphophonological learning begins: ① what sounds are possible in the language, and ② what permutations of meanings are possible in the language. In order to elaborate on these concepts, I will discuss each in their own sections below.

2.1.1 What sounds are possible in the language?

I follow common practice and assume that the model knows in advance which sounds are possible in the language. I pre-define the sounds of the language via an inventory of

Table 2-1: Sample segment and feature inventory.

	[cons]	[cor]	[voi]
t	1	1	-1
d	1	1	1
k	1	-1	-1
g	1	-1	1
a	-1	-1	1

SEGMENTS and FEATURES. A segment is a symbolic representation of phonological features, the latter of which can take the value of ±1 or 0. A sample inventory is given in Table 2-1. The inventory specifies which sounds are possible; segments not specified in the inventory are assumed to not exist in the language.

The concatenation of one or more segments in a sequence results in a FORM; forms that are stored in the lexicon are known as the UNDERLYING FORMS U, or URs, whereas forms that are observed by the model in training are known as the OBSERVED FORMS O, or ORs, of the language. I apply no constraint on the space of possible forms that the inventory can generate; it is up to the phonology in order to determine which forms are licit in the language (cf. RICHNESS OF THE BASE, Smolensky 1996).

2.1.2 What meaning permutations are possible in the language?

I follow previous models of joint lexicon-grammar learning and assume that the learner is aware of which atomic meanings exist in the language, as well as which permutations these meanings are allowed to exist in. The space of valid meaning permutations is defined *a priori*, and any permutations absent in the defined space is assumed by the model to not be possible. I refer to the set of atomic meanings as the LEXEMES X of the language. Moving forward, I will refer to specific lexemes using the notation μ_x, where $x \in X$. I moreover refer to the space of possible valid lexeme permutations as the LEXICAL CONTEXTS C of the language. Moving forward, I will refer to specific lexical contexts using the notation $\langle \mu_c \rangle$, where $c \in C$. As a lexical context is comprised of several individual lexemes, specific lexemes found in a given lexical context will be identified using the notation μ_{xc}.

2.1.3 Working through a sample language: final-devoicing

In order to better elucidate the concepts introduced above, consider the sample language in Table 2-2. The ORs found in the language consist only of the segments specified in the sample inventory in Table 2-1. The language is organized into two sub-tables: the

Table 2-2: Sample language demonstrating final-devoicing.

Alternating				Non-alternating			
tak	$\langle \mu_1 \rangle$	taga	$\langle \mu_1, \mu_A \rangle$	tak	$\langle \mu_5 \rangle$	taka	$\langle \mu_5, \mu_A \rangle$
tat	$\langle \mu_2 \rangle$	tada	$\langle \mu_2, \mu_A \rangle$	tat	$\langle \mu_6 \rangle$	tata	$\langle \mu_6, \mu_A \rangle$
???	$\langle \mu_3 \rangle$	kaga	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_7 \rangle$	kaka	$\langle \mu_7, \mu_A \rangle$
???	$\langle \mu_4 \rangle$	kada	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_8 \rangle$	kata	$\langle \mu_8, \mu_A \rangle$

alternating and non-alternating stems. Each row in a sub-table designates a PARADIGM, or the space of lexical contexts a given lexeme is found in. For example, the paradigm for the lexeme μ_1 consists of two lexical contexts: the lexeme μ_1 ① in isolation, denoted by $\langle \mu_1 \rangle$, or ② followed by μ_A, denoted by $\langle \mu_1, \mu_A \rangle$. Note that not all permutations are possible in this language; lexical contexts such as $\langle \mu_1, \mu_2 \rangle$ and $\langle \mu_A, \mu_1 \rangle$ are noticeably absent.

Associated with each lexical context is either an OR or unobserved form. For example, the lexical context $\langle \mu_3, \mu_A \rangle$ is associated with the OR [kaga], while the lexical context $\langle \mu_3 \rangle$ has no observed counterpart, indicated using "???". Moving forward, I will use the term paradigm to refer to both the space of lexical contexts for a given lexeme as well as their associated ORs. Paradigms that consist of all ORs are known as COMPLETE PARADIGMS, while paradigms that contain some unobserved forms are known as INCOMPLETE PARADIGMS.

The sample language given in Table 2-2 consists of several complete alternating and non-alternating paradigms, demarcated by the non-shaded cells. The alternating paradigms correspond to the lexical contexts containing either the lexeme μ_1 or the lexeme μ_2, and are labelled as such due to the alternation observed between the voiceless and voiced stops T~D. Which stop surfaces depends on the context: the stop surfaces as voiced when found intervocalically and voiceless when found word-finally: [taga] and [tak] for the lexical contexts $\langle \mu_1, \mu_A \rangle$ and $\langle \mu_1 \rangle$, respectively. In contrast, the non-alternating paradigms correspond to the lexical contexts containing μ_5 and μ_6. These paradigms retain the voiceless stop regardless of context: [taka] and [tak] for the lexical contexts $\langle \mu_5, \mu_A \rangle$ and $\langle \mu_5 \rangle$, respectively.

The ORs in the data were generated via the application of a single phonological process of final-devoicing: O → T / _#. Stems in the alternating paradigms consist of URs with stem-final voiced consonants. For example, the URs for the lexical contexts $\langle \mu_1, \mu_A \rangle$ and $\langle \mu_1 \rangle$ are /tag-a/ and /tag/, respectively. Stems in the non-alternating paradigms, in contrast, consist of URs with stem-final voiceless consonants. For example, the URs for the lexical contexts $\langle \mu_5, \mu_A \rangle$ and $\langle \mu_5 \rangle$ are /tak-a/ and /tak/, respectively. Note that the ORs provided are UNSEGMENTED, or do not give any information as to which segments are

associated with which lexeme; it is up to the model to determine how to distribute the OR into each of the lexemes within that lexical context, e.g. ascertain that the OR [taka] for the lexical context $\langle \mu_5, \mu_A \rangle$ was generated from the UR /tak-a/. Applying the final-devoicing process to the URs produces their respective ORs, as demonstrated in (1).

(1) Lexicon and grammar used to generate the ORs of the sample language. From left to right, the lexical contexts for each UR are: $\langle \mu_1, \mu_A \rangle$, $\langle \mu_1 \rangle$, $\langle \mu_5, \mu_A \rangle$, and $\langle \mu_5 \rangle$

Underlying Form	/tag-a/	/tag/	/tak-a/	/tak/
O → T / _#	–	[tak]	–	–
Observed Form	[taga]	[tak]	[taka]	[tak]

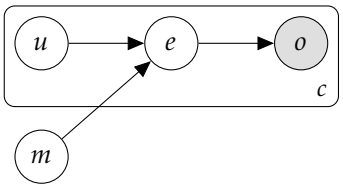
Observe that the ORs of some paradigm slots between the alternating and non-alternating stems are identical; for example, the lexical contexts $\langle \mu_1 \rangle$ and $\langle \mu_5 \rangle$ both consist of the OR [tak], despite the former belonging to the alternating category and the latter belonging to the non-alternating category. This distribution is intended to provide information to the learner that the language was generated through final-devoicing as opposed to an alternative analysis, such as one involving intervocalic voicing: O → D / V_V. The intervocalic voicing analysis would fail to explain why the form [taka] does not voice [k] to [g].

In addition to the complete paradigms, the sample language also consists of several incomplete paradigms for the lexemes μ_3 to μ_4, as well as μ_7 to μ_8, marked by the shaded cells. These paradigms are comprised of ORs for the complex lexical contexts, e.g. $\langle \mu_3, \mu_A \rangle$ and $\langle \mu_7, \mu_A \rangle$, as well as the unobserved forms for the simplex contexts, e.g. $\langle \mu_3 \rangle$ and $\langle \mu_7 \rangle$. We refer to the latter forms as the held-out forms. The goal of the model is to infer which lexicon-grammar pairs fit well with the given lexical contexts and ORs. This distribution will then be used to predict the forms for each of the held-out lexical contexts.

2.2 The Generative Model

The basic architecture of the generative lexicophonological model is provided in Figure 2-1. This model consists of two components: ① the lexicon, or the set of URs for each lexeme u_x, and ② the grammar, or the phonological mappings m that transform the UR into its eventual OR. The model assumes that the ORs are generated in three steps. ① The model builds the UR for each lexeme u_x. These URs are concatenated to construct the URs for each lexical context $u_c = u_{x_1 c} + \cdots + u_{x_k c}$. ② To each UR, the model selects and applies a phonological mapping m to generate an expected form $e_c = m(u_c)$, hereafter, ER. ③ The

Figure 2-1: Basic architecture of the generative lexicophonological model.



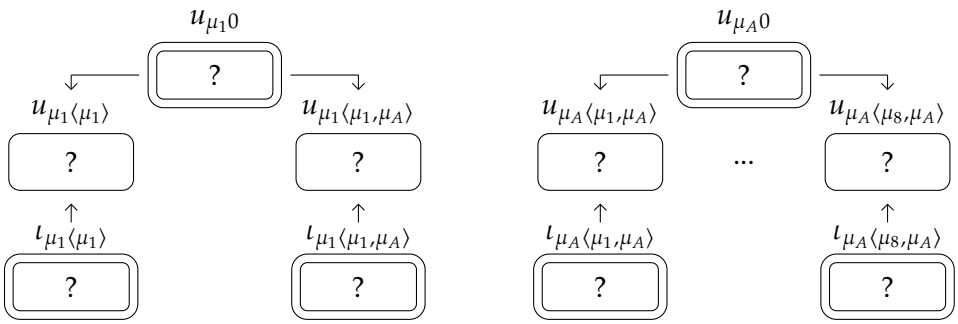
model applies noise to each ER, potentially corrupting and transforming them into the ORs of each observed lexical context o_c. Note that unlike in the previous section, in which the ORs were generated from the grammar directly, the model formulation here instead incorporates an extra step in which the output of the grammar is corrupted by noise before being observed by the learner.

In order to better demonstrate the components of the generative model, I will be using the sample inventory and language given in Table 2-1 and Table 2-2 above as a running example. I walk through and explain each step in more detail in the following sections.

2.2.1 Generating the underlying forms

I assume a generative model over URs for each lexeme u_x. A UR for a lexeme is generated in two stages. ① The model generates a PROTOTYPE UNDERLYING FORM u_{x0}. The prototype UR corresponds the lexeme's abstract DEFAULT form; it dictates the general shape all URs belonging to the lexeme should have. ② For each lexical context the given lexeme is found in, the model generates a potentially distinct, context-specific UR known as the CONTEXTUAL UNDERLYING FORM u_{xc}; it is the actual UR fed into and manipulated by the grammar in order to produce the ERs. The method of generating each contextual UR is dictated by an IDENTITY PARAMETER ι_{xc}, which determines the shape of the contextual UR either by directly reusing the prototype UR or by generating a new UR specific to that lexical context. Schematizations of the structures of the URs for the lexemes μ_1 and μ_A are provided in (2).

(2) Schematic UR structures for the lexemes μ_1 and μ_A



As can be observed, the UR structure for the lexeme μ_1 consists of the prototype UR, labeled $u_{\mu_1 0}$, and contextual URs for the following two lexical contexts: the lexeme μ_1 ① in isolation, labeled $\langle \mu_1 \rangle$, and ② followed by the lexeme μ_A, labeled $\langle \mu_1, \mu_A \rangle$. Likewise, the lexeme μ_A consists of the prototype UR, labeled $u_{\mu_A 0}$, and eight contextual URs, each associated with the lexical contexts in which the lexeme μ_A is preceded by the lexemes μ_1 to μ_8. I will continue to work with this schematic structure as we explain the generative processes for the prototype and contextual URs in the sections below.

Generating the prototype underlying forms u_{x0}

Given a segment inventory of size $|S|$, the prototype UR is generated by sampling w segments from a distribution proportional to the following unnormalized probability mass function in (3). This distribution takes a single hyperparameter, $0 \leq \theta$, which dictates the general length of the prototype UR.

(3) $\quad u_{x0} \sim |S|^{-\theta w}$

The sample inventory in Table 2-1 consists of 5 segments {t, k, g, i, a}, resulting in the following updated distribution:

(4) $\quad u_{x0} \sim (5)^{-\theta w}$

How does the hyperparameter θ affect the length of the generated prototype UR? I plot the truncated distribution over possible prototype UR lengths $0 \leq w \leq 10$ for three different values of θ in Figure 2-2.[1] We see that when $\theta = 0$, all prototype UR lengths are equally probable; however, as we raise the value of θ, the probability mass increasingly shifts towards 0. In other words, under this distribution, the larger θ is, the more biased the model will be towards shorter URs. For the discussion here, I set $\theta = 0.5$.

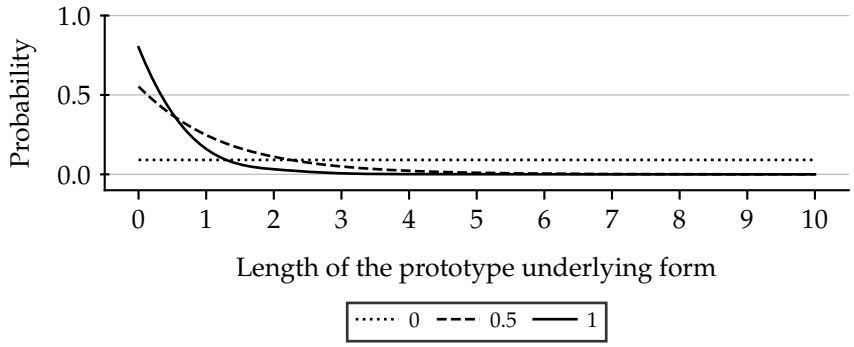
To see this in practice, let us say the model generates for $u_{\mu_1 0}$ and $u_{\mu_A 0}$ the prototype URs /ta/ and /ka/, respectively. We can compute the unnormalized probability of generating each form, shown in (5). Observe that URs of the same length have identical probabilities assigned to them: $P(/\text{ta}/) = P(/\text{ka}/) = P(/\text{aa}/) = P(/\text{tt}/) \ldots \propto 0.2$.[2]

(5) $\quad P(/\text{ta}/) \propto (5)^{-0.5 \times 2} \propto 0.2 \qquad P(/\text{ka}/) \propto (5)^{-0.5 \times 2} \propto 0.2$

[1] Due to performance reasons, I truncate the distributation by applying an upper bound. In all of the simulations moving forward, we fix the upper bound to be 5: $0 \leq |u_{x0}| \leq 5$.

[2] I do not apply any restriction on the space of URs; for example, URs that consist only of consonants such as /tt/ are considered just as probable as forms like /ta/. I leave exploring how adjusting the space of URs affects the model's predictions to future work.

Figure 2-2: Normalized probabilities over prototype UR lengths for $\theta \in \{0, 0.5, 1\}$.

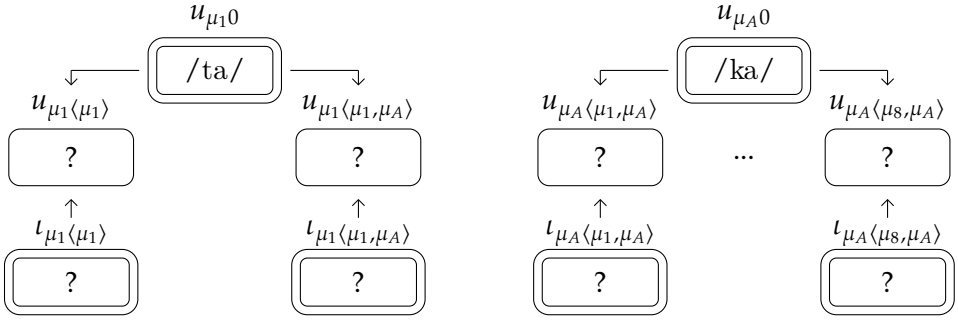


Contrast this with the alternate set of prototype URs /tag/ and /a/. We see in (6) that the longer prototype UR /tag/ has a comparatively lower probability than its shorter counterpart /ta/, i.e. 0.09 versus 0.2, respectively.

(6) $P(/\text{tag}/) \propto (5)^{-0.5\times3} \propto 0.09 \qquad P(/\text{a}/) \propto (5)^{-0.5\times1} \propto 0.45$

At the end of this stage, let us suppose that the model generates the prototype URs /ta/ and /ka/ for $u_{\mu_1 0}$ and $u_{\mu_A 0}$. This nets us the updated structure in (7).

(7) Updated schematic UR structure for the lexemes μ_1 and μ_A



Generating the contextual underlying forms u_{xc}

For each lexical context a lexeme is found in, the model must generate a contextual UR. Each contextual UR is generated in two steps. In the first step, the model samples an identity value to determine whether to use the existing prototype UR or to generate a new contextual UR ι_{xc}. The identity value is drawn from a Bernoulli distribution, shown in (8). This distribution takes a single parameter, the probability of success $0 \leq \alpha \leq 1$, and returns a binary value indicating whether to use the prototype UR or to construct a new UR for that lexical context.

(8) $\iota_{xc} \sim \text{Bernoulli}(\alpha)$

In our working example, suppose we set $\alpha = 0.75$. We update the distribution accordingly:

(9) $\iota_{xc} \sim \text{Bernoulli}(0.75)$

Under this parameterization, the probability of the model choosing to use the default prototype UR is 75%, while the probability of the model generating a new contextual UR is 25%. Thus, the model will favor reusing the prototype UR rather than proposing a new, unique UR for each lexical context. The model is biased towards capturing alternation through the productive application of phonological processes rather than by encoding the alternation into the UR. The ability to posit unique URs for lexemes in a specific context is reminiscent of the notion of CONTEXTUAL ALLOMORPHY. This allows the model to consider alternative methods of encoding alternations into the lexicon rather than producing them via the grammar (Paster 2005), as well as completely unique forms that cannot be predicted from the the phonology alone, i.e. *go* ~ *went* in English.

In the second step, depending on which identity value is sampled, the model performs a different action. If the model decides to use the prototype UR $\iota_{xc} = 1$, then the model simply reuses the prototype UR as its contextual UR for that lexical context.

(10) $u_{xc} \mid \iota_{xc} = 1, u_{x0} \sim \mathbb{1}\left[u_{xc} = u_{x0}\right]$

In contrast, if the model does not decide to use the prototype UR $\iota_{xc} = 0$, then a new contextual UR must be generated. This is done by sampling a string from a multinomial distribution based on the minimum edit distance between it and the prototype UR $d_{u_{x0}}$. This distribution takes a single parameter $0 \leq \psi$, which dictates the level faithfulness of the contextual UR to the prototype UR.[3]

(11) $u_{xc} \mid \iota_{xc} = 0, u_{x0} \sim \text{Multinomial}(-\psi d_{u_{x0}})$

The minimum edit distance is computed by calculating the minimum number of possible weighted edits to get from one string to another string.[4] There are many possible string edit measures to choose from, but the specific metric I adopt is the WEIGHTED LEVENSHTEIN DISTANCE, which defines three possible edit operations over strings: insertions, deletions, and substitutions. For each operation, we assign a cost of applying that transformation.

[3] As above for the prototype URs, I fix the upper bound of the length of contextual URs to be 5: $0 \leq |u_{xc}| \leq 5$.

[4] The sequence of edit operations with the lowest edit distance to get from the source to target string is known as the OPTIMAL ALIGNMENT. See Wagner and Fischer (1974) for details on how the optimal alignment is determined.

For each possible pair of input segments x and output segments y between the prototype UR and possible contextual UR, I adopt the cost function f in (12). Under this cost function, for insertions and deletions, we assign an arbitrary cost of 1. For substitutions, I vary the cost as the complement of the similarity between the two segments; the more similar the two segments are, the lower the cost is to substitute one for the other.

(12) Proposed cost function f for each edit operation

a. Insertion: $f(*, y) = 1$

b. Deletion: $f(x, *) = 1$

c. Substitution: $f(x, y) = 1 - sim(x, y)$

I define the similarity of two segments using the metric laid out in Frisch et al. (2004). Here, similarity is determined by the relative proportion of shared natural classes between the two segments. This is typified in (13).

(13) $$sim(x, y) = \frac{\text{Shared natural classes}}{\text{Shared natural classes} + \text{Non-shared natural classes}}$$

The inventory in Table 2-1 consists of five segments {t, k, g, i, a} and three phonological features.[5] Given this inventory, I compute the similarity of several segment pairs in (14).

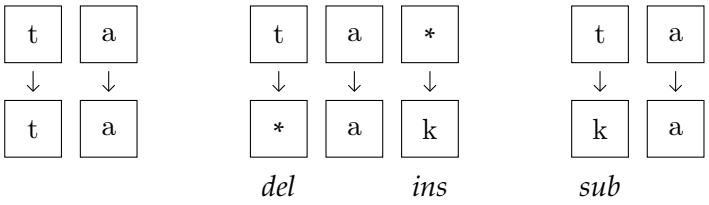
(14) $$sim(\text{t}, \text{t}) = \frac{7}{7+0} = 1 \qquad sim(\text{t}, \text{k}) = \frac{3}{3+4} = 0.43$$
$$sim(\text{t}, \text{g}) = \frac{1}{6+1} = 0.14 \qquad sim(\text{t}, \text{a}) = \frac{0}{0+7} = 0$$

Identical segments have a similarity of 1 as they share the exact same natural classes. As the segments become increasingly featurally contrastive, however, the similarity score drops and the edit cost increases. For example, the pair [t] and [k] share in voicing and consonantality, but differ in place, while the pair [t] and [g] share only in consonantality, differing in both voicing and place. The aforementioned pairs consequently have proportionally lower similarity scores: 0.43 and 0.14, respectively. Finally, the pair [t] and [a] shares no overlapping natural classes and thus has a similarity of 0.

Combining all these concepts together, let us calculate the minimum edit distance between the sampled prototype UR /ta/ and three possible contextual URs /ta/, /ak/, and /ka/. Under the proposed cost function, the minimum edit distance between the strings /ta/ and /ta/ is 0 as there are no non-identity relations in the optimal alignment.

[5] The original design of the similarity metric included an additional feature [segment] such that all pairs of segments had some kind of featural overlap. I omit this feature in order to allow for some substitutions to be as costly as insertions or deletions, i.e. have a cost of 1.

Figure 2-3: Alignments from the prototype UR /ta/ to the contextual URs /ta/, /ak/, and /ka/.

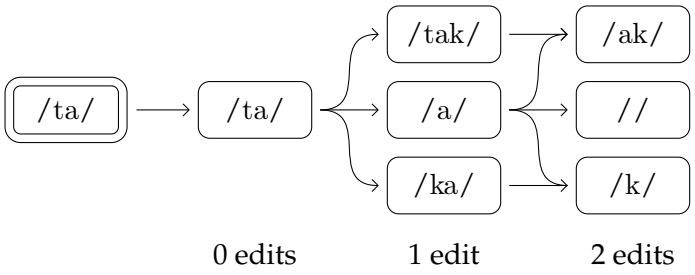


In constrast, the minimum edit distance between the strings /ta/ and /ak/ is 2 due to there being both a deletion and insertion operation needed in the optimal alignment. Lastly, the minimum edit distance between the strings /ta/ and /ka/ is 0.57, as the cost of substituting [t] to [k] under the optimal alignment is the complement of their similarity: $0.57 = 1 - 0.43$. The edit operations under the optimal alignments are shown in Figure 2-3.

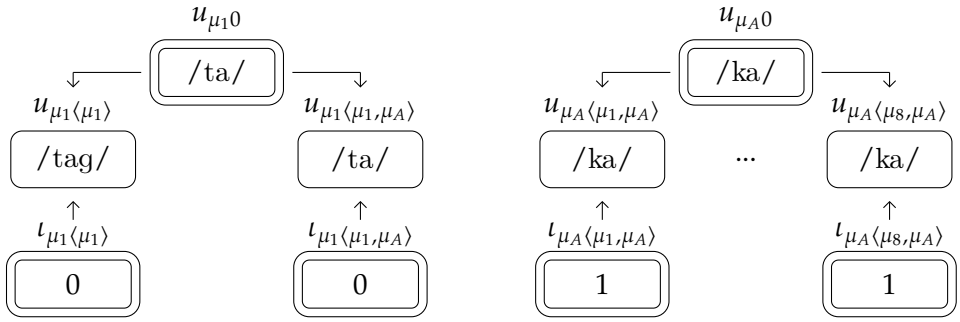
How does the edit distance metric affect which contextual URs are sampled? I map out in Figure 2-4 a subset of the space of possible contextual URs for the lexical context $\langle \mu_1 \rangle$ given the prototype UR for the lexeme μ_1 /ta/. As can be observed, the more edits that are applied to the initial input string, the more different the resulting string is. Thus, the edit distance penalizes contextual URs that stray too far from its prototype UR. The larger ψ is, the more similar the generated contextual UR will be to the prototype UR. Much like with α, a high ψ favors contextual URs that do not vary too much from the default UR. For the discussion here, I set $\psi = 5$. Note that due to the way the distribution is formalized, the model may generate a contextual UR that is identical in form to the prototype UR.

The lexeme μ_1 in the toy language is found in two lexical contexts: ① in isolation, denoted by $\langle \mu_1 \rangle$, and ② followed by the lexeme μ_A, denoted by $\langle \mu_1, \mu_A \rangle$. The model therefore must generate two contextual URs: ① $u_{\mu_1\langle\mu_1\rangle}$ and ② $u_{\mu_1\langle\mu_1,\mu_A\rangle}$. Suppose that the model samples $\iota_{\mu_1 c} = 0$ for both lexical contexts μ_1 is found in. As a result, the model must generate a new contextual UR for each context. Since the parameterization favors high faithfulness to the default prototype UR, the model will generate contextual URs similar to the already sampled prototype UR /ta/. Let us suppose the model samples the contextual URs /tag/ and /ta/ for the lexical contexts $\langle \mu_1 \rangle$ and $\langle \mu_1, \mu_A \rangle$ for lexeme μ_1, respectively. In contrast, the lexeme μ_A is found in eight lexical contexts, following each of the lexemes μ_1 to μ_8. This means the model must generate eight contextual URs for this lexeme. Let us suppose that the model happens to sample $\iota_{\mu_A c} = 1$ for all contexts μ_A is found in. This results in the model reusing the prototype UR for each contextual UR rather than generating one anew. As the prototype UR for lexeme μ_A is /ka/, the model consequently updates all contextual URs for the lexeme to be /ka/. This nets us the completed UR structure shown in (15).

Figure 2-4: Subspace of contextual URs from the prototype UR /ta/ by number of edits.



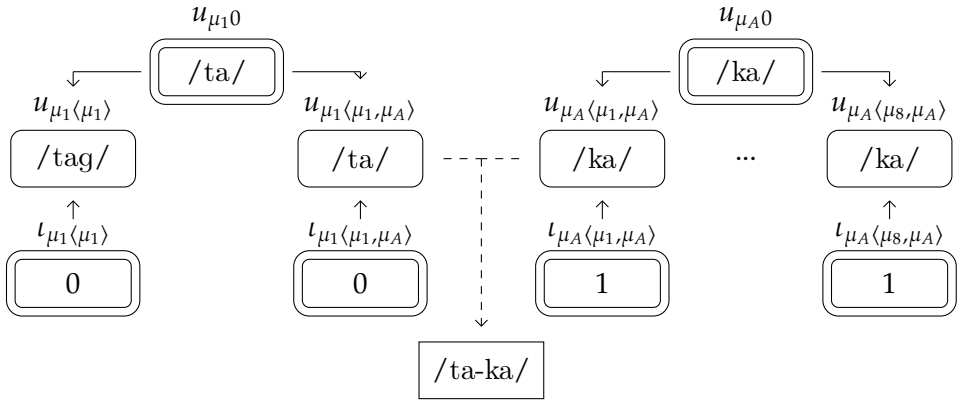
(15) Completed schematic UR structure for lexemes μ_1 and μ_A



Generating the underlying form for the lexical contexts u_c

The UR for a lexical context is the concatenation of each of the contextual URs for the lexemes in the lexical context. For example, the UR for the lexical context $\langle\mu_1, \mu_A\rangle$ is generated by concatenating the contextual URs $u_{\mu_1\langle\mu_1,\mu_A\rangle}$ and $u_{\mu_A\langle\mu_1,\mu_A\rangle}$. Given the current contextual UR hypotheses for each lexeme, the model would predict the UR of /ta-ka/. This is visualized in (16).

(16) Generating the UR for the lexical context $\langle\mu_1, \mu_A\rangle$



2.2.2 Generating the expected forms e_c

Once the model generates the prototype and contextual URs for each lexeme and the lexical contexts they appear in, the model produces the ER by applying the current mapping hypothesis $m \in M$ to the concatenated URs of each lexical context:

(17) $e_c = m(u_c)$

I define a mapping as any phonological transformation that potentially transforms a given UR of a lexical context and produces an ER. Any phonological theory satisfies this requirement, but I will focus our attention on a rule-based (i.e. SPE) formalization here. Since the main focus of the dissertation is on the learnability of process interactions, I choose to adopt a phonological system with no inherent bias towards one process interaction over another. Rule-based approaches carry this exact property: one ordering of two independently motivated rules is just as *a priori* likely is its opposite ordering. This is not true in most constraint-based approaches, which favor transparent interactions over their opaque counterparts.

The space of possible rule hypotheses M is the set of partially-ordered cartesian products of all possible non-repeating sub-sequences of the set of atomic rules in a given space. Let us specify the set of atomic rules to consist of two phonological processes: an intervocalic voicing process and a final-devoicing process. The rule definitions are defined in (18).[6]

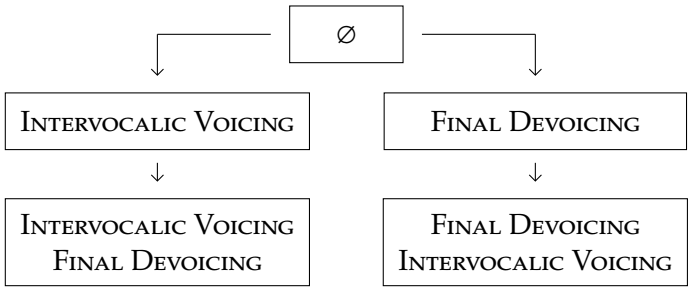
(18) a. Intervocalic Voicing: O → D / V_V
b. Final Devoicing: O → T / _#

The space of possible rule hypotheses M for this space of atomic rules is given in Figure 2-5. This space includes both the hypothesis in which there are no phonological rules as well as each possible ordering of the two atomic rules.

Given the UR hypotheses generated in the previous section for lexemes μ_1 and μ_A above, the model predicts the URs /tag/ and /ta-ka/ for the lexical contexts $\langle \mu_1 \rangle$ and $\langle \mu_1, \mu_A \rangle$, respectively. Let us suppose that the model applies final-devoicing to each UR. This results in the expected outputs given in (19).

[6] The space of atomic rules presented here – and thus the space of rule hypotheses – is completely arbitrary. I could have proposed other atomic rules such as a segment-specific final-devoicing process: g → k / _#. Intuitively, altering the space of hypotheses affects the predictions of the model in turn. I discuss the potential impact of changing these assumptions in Chapter 5, but leave exploring how changing the hypothesis space over rules affects model performance to future research.

Figure 2-5: Sample space of possible rule hypotheses M.



(19) Expected outputs for the lexical contexts $\langle \mu_1 \rangle$ and $\langle \mu_1, \mu_A \rangle$ given final-devoicing

Underlying Form	/tag/	/ta-ka/
O → T / _#	[tak]	–
Expected Form	[tak]	[taka]

I assume that the output of the grammar is deterministic; in other words, a phonological transformation will always occur if the structural description for its application is met.

2.2.3 Generating the observed forms o_c

So far, I have described how the model generates the lexicon and grammar and produces predictions or ERs for each lexical context; however, how do we relate the ERs of the model to the actual ORs given in the data? For readability, I repeat the sample language introduced in Table 2-2. The generated lexicon and grammar predicts the ERs [tak] and [taka] for the lexical contexts $\langle \mu_1 \rangle$ and $\langle \mu_1, \mu_A \rangle$, respectively. However, the actual ORs for each lexical context are slightly different, being [tak] and [taga], respectively.

I make the claim that learners, when given a set of ORs, carry a level of uncertainty in the reliability of the data and as consequence the hypotheses used to generate the data. The idea that learners' behavior reflects the combined influence of many different hypotheses is consistent with a broad class of models known as noisy channel models (Feldman & Griffiths 2007; Levy 2008), which have previously been applied in the acquisition literature to syntactic learning (Perkins, Feldman, & Lidz 2017; Schneider, Perkins, & Feldman 2020). These models assume that the data learners observe may have been corrupted by a noise process, which creates additional uncertainty about what the uncorrupted data look like. I distinguish between ERs and ORs along this dimension: the ERs correspond to the forms generated by the lexicon and grammar, whereas the ORs correspond to the forms observed

Table 2-2: Sample language demonstrating final-devoicing.

Alternating				Non-alternating			
tak	$\langle \mu_1 \rangle$	taga	$\langle \mu_1, \mu_A \rangle$	tak	$\langle \mu_5 \rangle$	taka	$\langle \mu_5, \mu_A \rangle$
tat	$\langle \mu_2 \rangle$	tada	$\langle \mu_2, \mu_A \rangle$	tat	$\langle \mu_6 \rangle$	tata	$\langle \mu_6, \mu_A \rangle$
???	$\langle \mu_3 \rangle$	kaga	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_7 \rangle$	kaka	$\langle \mu_7, \mu_A \rangle$
???	$\langle \mu_4 \rangle$	kada	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_8 \rangle$	kata	$\langle \mu_8, \mu_A \rangle$

by the learner. This forces the learner to consider not only hypotheses that are consistent with the data, but also hypotheses that are almost consistent with the data.

The implementation of the noisy channel I adopt is one used in Levy (2008), formalized in (20). Under this implementation, an OR is assumed to be generated from a distribution based on the weighted Levenshtein distance between it and the predicted ER. This distribution takes a single parameter $0 \leq \lambda$, which dictates the amount of noise in the model. I will assume the same cost function stipulated in (12), where insertions and deletions have a cost of 1 and substitutions between segments have a cost corresponding to the complement of their similarity.

(20) $o_c \mid e_c \sim \textsc{Multinomial}(-\lambda d_{e_c})$

We can compute the weighted edit costs to transform the ERs [tak] and [taka] into their respective ORs [tak] and [taga] in (21).

(21) $e_{\langle \mu_1 \rangle} = [\text{tak}] \rightsquigarrow [\text{tak}] = o_{\langle \mu_1 \rangle}$ (weighted edit cost = 0)

$e_{\langle \mu_1, \mu_A \rangle} = [\text{taka}] \rightsquigarrow [\text{taga}] = o_{\langle \mu_1, \mu_A \rangle}$ (weighted edit cost = 0.43)

As we can see, more edits correspond to a noisier output. The hyperparameter λ thus corresponds to a sensitivity term: the smaller the λ, the more noisy the output is believed to be, and consequently the more lenient we are with the hypotheses we consider. The noisy channel therefore allows the model to consider joint lexicons and grammars that do not perfectly match the data, but prefer those that do.

2.2.4 Summary of the generative procedure

In the previous sections, we went over how the model generates lexicons and grammars, and how these are in turn used to create the ERs and eventually the ORs of the language. I summarize the entire generative procedure in (22).

(22) a. For each lexeme, generate the prototype UR $u_{x0} \sim |S|^{-\theta w}$
b. For each lexical context the lexeme is found in, generate the contextual UR u_{xc}
(i) Determine the identity value $\iota_{xc} \sim \textsc{Bernoulli}(\alpha)$
(ii) If $\iota_{xc} = 1$: $u_{xc} \mid \iota = 1, u_{x0} \sim \mathbb{1}[u_{xc} = u_{x0}]$
(iii) If $\iota_{xc} = 0$: $u_{xc} \mid \iota = 0, u_{x0} \sim \textsc{Multinomial}(-\psi d_{u_{x0}})$
c. Concatenate the associated contextual URs $u_c = u_{x_1 c} + \cdots + u_{x_k c}$
d. For each generated UR, apply the mapping to produce the ER $e_c = m(u_c)$
e. For each ER, apply noise to generate the OR $o_c \mid e_c \sim \textsc{Multinomial}(-\lambda d_{e_c})$

2.3 Performing Inference Over the Lexicon and Grammar

In order to determine which lexicon-grammar pairs in this hypothesis space were most likely to have generated the ORs of the data, we examine the posterior distribution $P(h \mid d)$, which allows us to quantify this relationship by assigning a probability to a hypothesis h given the set of observed data d. In order to calculate and interpret this distribution, we use Bayes' theorem, giving us the identity in (23).

$$P(h \mid d) = \frac{P(d \mid h)\, P(h)}{P(d)} \tag{23}$$

The equation states that the probability assigned to a hypothesis given the observed data is equivalent to the likelihood assigned to the data given the hypothesis $P(d \mid h)$, or how well the predictions of the hypothesis match the observed data, times its prior $P(h)$, or the beliefs of the model regarding which hypotheses were used to generate the data before any data is observed. In our case, the hypotheses correspond to joint lexicon-grammar pairs, and the data corresponds to the OR-meaning pairs. We can rewrite the variables in (23) with ones we are familiar with, getting us the revised equation in (24).

$$P(m, u \mid o) = \frac{P(o \mid m, u)\, P(m, u)}{P(o)} \tag{24}$$

The likelihood and prior probabilities are normalized by the prior probability of the data, but as the set of data is fixed, we can drop the denominator and calculate just the numerator:

$$P(m, u \mid o) \propto P(o \mid m, u)\, P(m, u) \tag{25}$$

Now the joint posterior probability of the lexicon and grammar hypothesis given the observed data is proportional to the likelihood of the data given the hypothesis times its prior. I have not yet defined a distribution that gives us the likelihood directly. However, I

have defined a distribution that gets us from the lexicon and grammar to the ERs $P(e \mid m, u)$ and a distribution that gets us from the ERs to the ORs $P(o \mid e)$:

(26) $P(m, u \mid o) \propto \sum_e P(o \mid e)\, P(e \mid m, u)\, P(m, u)$

Since each joint lexicon and grammar produces categorical outputs, we can simplify the equation to get the distribution in (27).

(27) $P(m, u \mid o) \propto P(o \mid m, u)\, P(m, u)$
where $P(o \mid m, u) = \prod_c P(o_c \mid m, u_c) = \prod_c P(o_c \mid e_c = m(u_c))$
where $P(m, u) = P(m)\, P(u) = P(m) \prod_x P(u_x)$

We now have an interpretable joint posterior distribution: the joint probability of a lexicon and grammar hypothesis given the data is proportional to the product of the likelihood of the ORs given the ERs of the lexicon and grammar hypothesis times the prior of the hypothesis. The likelihood $P(o \mid m, u)$ in our case refers to the noisy channel component of our model, which quantifies how similar the ERs predicted by the lexicon and grammar match the ORs of the data; ERs with low weighted Levenshtein distance conversely have higher likelihood. The prior over the URs $P(u)$ in turn corresponds to the generative process over the lexicon in the model. As discussed above, this process encodes a preference for a lexicon in which the prototype UR is reused as much as possible, but still allows the model to consider contextual URs that vary according to the lexical context its associated lexeme is found in. I did not define any preference over the space of possible mappings $P(m)$; *a priori*, all rule hypotheses are considered equally probable: $P(m) \propto 1$.

While I have so far enumerated how the model computes the joint posterior for a single lexicon-grammar hypothesis, what I am actually interested in is knowing the *space* of hypotheses a group of learners may have come up with given the same set of data; as such, we want to calculate the entire joint posterior. In order to do so, however, I would need to calculate the probabilities of all lexicon-grammar hypotheses. Given that I am working with a generative model over URs, the hypothesis space is prohibitively large to enumerate. To circumvent this issue, I will approximate the space instead. I will be using a class of algorithms designed to approximate these distributions known as Markov Chain Monte Carlo (MCMC) methods. Specifically, I will employ two algorithms from this class, which estimate the posterior by sampling hypotheses from the posterior in slightly different ways. I go over each of the algorithms in the subsections below.

2.3.1 The Gibbs sampling algorithm

The first sampling algorithm I use is that of the GIBBS SAMPLING ALGORITHM (Geman & Geman 1984). The algorithm works by sequentially sampling and updating each parameter in the model given the current hypotheses of the other parameters. The space of parameters in the model correspond to the LEXICON, or the prototype URs for each lexeme u_{x0} and contextual URs and identity values for each lexeme in each lexical context u_{xc} and ι_{xc}, as well as the GRAMMAR, or the sequence of mappings m that apply to the URs of each lexical context. As before, in order to better illustrate how the Gibbs sampling algorithm works, I will walk through how inference proceeds using the final-devoicing sample language presented earlier.

Sampling the URs for each lexical context

First, the model samples the UR hypotheses for each lexical context. This process is broken down into two steps. In the first step, the model goes through and samples the URs and identity settings for each possible lexical context in the data, u_c and ι_c. In other words, we are sampling from the following conditional distribution:

(28) $P(u_c, \iota_c \mid u_{-c}, \iota_{-c}, u_0, m, o)$

The parameter settings for the URs of the other lexical contexts do not affect the joint probability, so we can drop them from the equation. This leaves us only the mapping hypothesis, as well as the prototype URs for each lexeme in the lexical context u_{x_c0}:

(29) $P(u_c, \iota_c \mid u_{x_c0}, m, o)$

Using the definition of conditional probabilities, joint probabilities, and the chain rule, we can then derive the identity in (30).

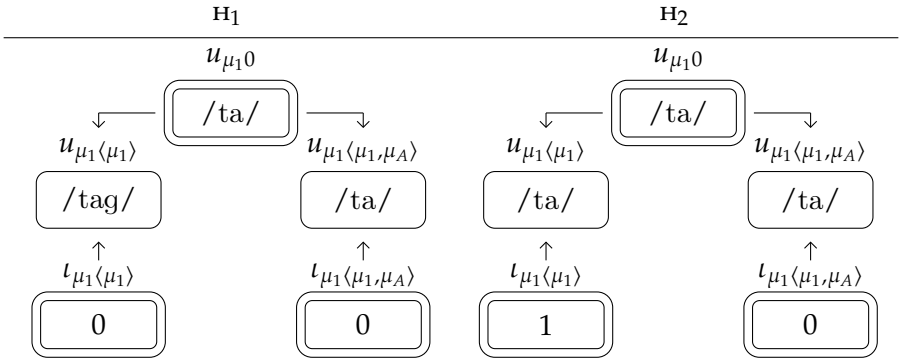
(30) $P(u_c, \iota_c \mid u_{x_c0}, m, o) \propto P(o_c \mid m, u_c) \prod_{x \in x_c} P(u_{xc} \mid u_{x0}, \iota_{xc}) P(\iota_{xc})$

The probability of the set of identity values and URs for a given lexical context is proportional to the likelihood of the data given the proposed UR and current mapping hypothesis times the prior of the contextual UR and identity value of each lexeme.

Let us suppose that the model is sampling a UR for the lexical context $\langle \mu_1 \rangle$. As the lexeme μ_1 is the only lexeme in the lexical context, the model only needs to generate the identity value and contextual UR for that lexeme alone, i.e. $\iota_{\mu_1\langle\mu_1\rangle}$ and $u_{\mu_1\langle\mu_1\rangle}$. The parameter settings given in the preceding sections produce (non-exhaustive) hypothesis

space given in (31). Note that the only differences between the hypotheses are the identity of $u_{\mu_1\langle\mu_1\rangle}$ and $\iota_{\mu_1\langle\mu_1\rangle}$.

(31) Non-exhaustive contextual UR hypothesis space for the lexical context $\langle\mu_1\rangle$



In order to sample from this distribution, the Gibbs sampling algorithm looks at and calculates the likelihood and prior of every possible UR for each lexeme in the lexical context. Given the space of UR hypotheses given above, we would compute the relative joint probabilities as in (32).

(32) $P(u_{\langle\mu_1\rangle}, \iota_{\langle\mu_1\rangle} = \text{H}_1 \mid \dots) \propto P([\text{tag}] \mid m, \text{tag})\, P(\text{tag} \mid \text{ta}, 0)\, P(0)$
$P(u_{\langle\mu_1\rangle}, \iota_{\langle\mu_1\rangle} = \text{H}_2 \mid \dots) \propto P([\text{tag}] \mid m, \text{ta}\ )\, P(\text{ta}\ \mid \text{ta}, 1)\, P(1)$

Once the probabilities are calculated, the algorithm samples and updates the model with one of the hypotheses based on this distribution. Let us suppose the model selects H_2. The model then moves on, sampling and updating the UR hypothesis for the next lexical context, e.g. $u_{\langle\mu_1,\mu_A\rangle}$. For complex lexical contexts like $u_{\langle\mu_1,\mu_A\rangle}$, the model jointly samples the identity values and associated contextual URs for all lexemes found in the lexical context, i.e. $\iota_{\mu_1\langle\mu_1,\mu_A\rangle}$ and $\iota_{\mu_A\langle\mu_1,\mu_A\rangle}$, as well as $u_{\mu_1\langle\mu_1,\mu_A\rangle}$ and $u_{\mu_A\langle\mu_1,\mu_A\rangle}$.

Sampling the prototype URs

In the second step, once all the parameter settings for the URs of each lexical context have been sampled and updated, the model goes through and samples the prototype URs for each lexeme u_{x0}. In other words, we are sampling from the following conditional distribution:

(33) $P(u_{x0} \mid u_{-x0}, u_c, \iota_c, m, o)$

The parameter settings for the other prototype URs do not affect the probability of the current prototype UR; only the contextual URs u_{c_x} and identity settings ι_{c_x} in which the

given lexeme is present matter. As such, we can simplify the equation to the following:

(34) $P(u_{x0} \mid u_{c_x}, \iota_{c_x}, m, o)$

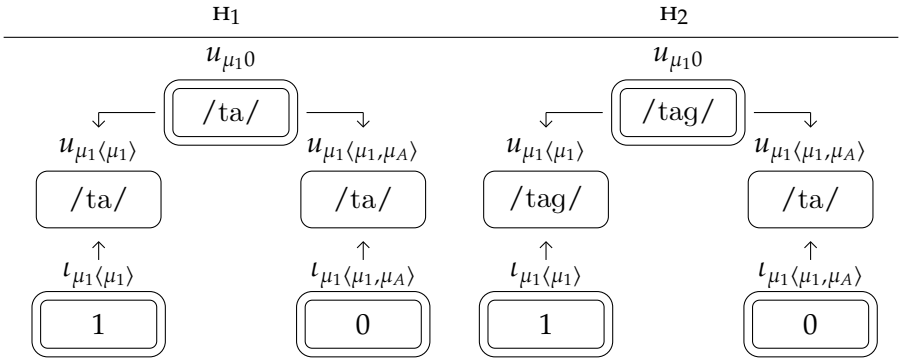
Again, using the definition of conditional probabilities and the chain rule, we derive the following identity:

(35) $P(u_{x0} \mid u_{c_x}, \iota_{c_x}, m, o) \propto \prod_{c \in c_x} P(o_c \mid m, u_c)\, P(u_{xc} \mid u_{x0}, \iota_{xc})\, P(u_{x0})$

The probability of a prototype UR is proportional to the likelihood of each lexical context the lexeme is found times the probability of each contextual UR of the lexeme as well its prior. Note that since some of the contextual URs reuse the prototype UR, the likelihood must be recalculated for each lexical context in which said default is utilized.

Let us say that the model is sampling a prototype UR for the lexeme μ_1. The hypothesis space consists of the following (non-exhaustive) set of hypotheses in (36). Observe that the contextual UR for the lexical context $\langle \mu_1 \rangle$ changes alongside the prototype UR.

(36) Non-exhaustive prototype UR hypothesis space for lexeme μ_1



In order to sample from this distribution, the Gibbs sampling algorithm looks at and calculates the probabilities of every possible prototype UR for that lexeme. Given the sample hypotheses above, we would compute the probabilities as shown in (37).

(37) $P(u_{\mu_1 0} = \text{H}_1 \mid \dots) \propto P([\text{tag}] \mid m, \text{ta}\ )\, P(\text{ta}\ \mid \text{ta},\ \ 1)\, P(\text{ta} \mid \text{ta},\ \ 0)\, P(\text{ta})$

$P(u_{\mu_1 0} = \text{H}_2 \mid \dots) \propto P([\text{tag}] \mid m, \text{tag})\, P(\text{tag} \mid \text{tag}, 1)\, P(\text{ta} \mid \text{tag}, 0)\, P(\text{tag})$

Once these probabilities are computed, the algorithm samples and updates the model with one of the hypotheses based on this distribution. The model then moves on, sampling and updating the prototype UR for the other lexemes.

Sampling the phonological mapping

Finally, once the UR hypotheses for each lexeme are updated, the model samples a mapping hypothesis given the data and newly sampled URs. The parameter settings for the prototype UR are irrelevant; thus, the right-hand side can be simplified to:

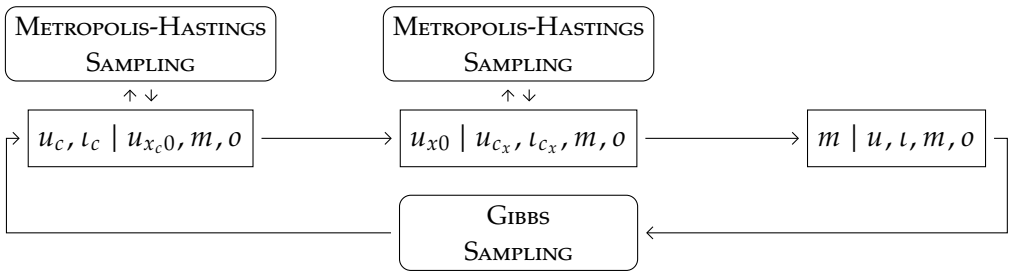
(38) $P(m \mid u, o) \propto \prod_c P(o_c \mid m, u_c)\, P(m)$

The algorithm calculates the likelihood of the entire dataset for each mapping hypothesis and, like before, samples a mapping hypothesis based on the joint likelihood and prior probability. Once a new mapping hypothesis is sampled, a single Gibbs iteration is completed. This process is repeated for g iterations, storing each fully-updated hypothesis at the end of each cycle. Once the algorithm finishes running, it removes the first half of the samples drawn as the initial samples are likely not indicative of the true posterior distribution. Moreover, as local samples are highly correlated, the algorithm only take every bth sample. This results in the final posterior sample.

2.3.2 The Metropolis-Hastings sampling algorithm

In the previous section, I went over how the model utilizes Gibbs sampling in order to approximate the joint posterior of the lexicon-grammar hypotheses. This is done by sampling from the conditional probability of each parameter given the values of the other parameters. However, since I am working with a generative model over URs, the hypothesis space is extremely large; as such, brute-force enumeration over all the URs in order to calculate and sample from these conditional probabilities is still not feasible. Instead, I must estimate the conditional probabilities for the contextual URs of each lexical context $P(u_c, \iota_c \mid u_{x_c 0}, m, o)$ as well as the prototype URs $P(u_{x0} \mid u_{c_x}, \iota_{c_x}, o)$ for each lexeme via the second algorithm: Metropolis-Hastings sampling, or more specifically, Metropolis-Hastings within Gibbs sampling (Hastings 1970). Within a Gibbs iteration, the Metropolis-Hastings sampling algorithm performs a separate pseudo-random walk over h iterations for each conditional distribution to be estimated. A single Metropolis-Hastings iteration follows two high-level steps. ① A hypothesis is proposed for the parameter. ② The new hypothesis is compared with the current hypothesis, which the model accepts or rejects based on their likelihoods, priors, and relative probability of being proposed. As we are working with two parameters and distributions – ① the URs and identity settings for each lexical context and ② the prototype URs of each lexeme – a separate Metropolis-Hastings process must be defined for each. I will first go over how the model estimates the distribution over contextual URs before moving on to the prototype URs.

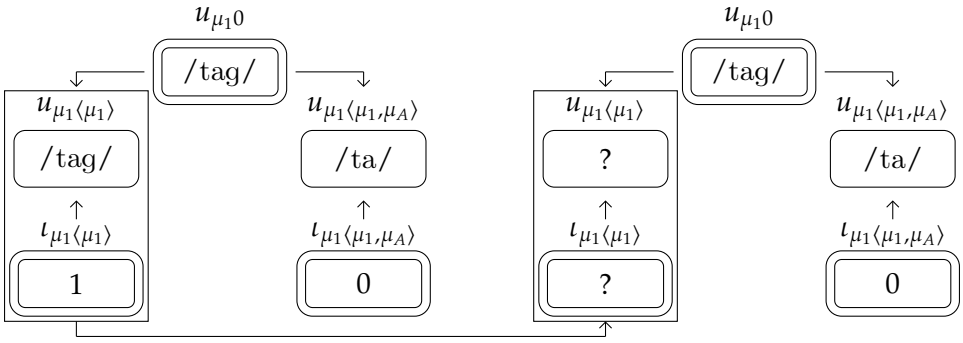
Figure 2-6: Visual schematization of the Metroplis-Hastings within Gibbs sampling algorithm. Within a single Gibbs iteration, the model performs a separate Metropolis-Hastings run for h iterations for each conditional distribution to be estimated.



Estimating the conditional probability over URs for a lexical context

In order to estimate the conditional distribution over the URs for each lexical context, the model performs a separate, independent Metropolis-Hastings run for each, including the held-out forms.[7] In order to illustrate this algorithm, we once again return to our running example; let us begin with a random UR hypothesis for μ_1, given in (39).

(39) Sampling procedure for the lexical context $\langle \mu_1 \rangle$



Suppose that the model is tryingto sample and update the UR hypothesis for $\langle \mu_1 \rangle$. As discussed with the Gibbs sampling algorithm, the model jointly samples two parameters: ① the identity value and ② the contextual UR for each lexeme in the lexical context. Since the identity value directly affects what contextual UR is sampled, the algorithm first samples a new identity value ι'_{xc}. The distribution it samples from is known as the PROPOSAL DISTRIBUTION Q, and is congruous with the probability of proposing some new parameter setting given the current parameter setting. In this case, it is the probability of either switching the identity value or keeping it the same $Q(\iota'_{xc} \mid \iota_{xc})$. This new identity parameter is drawn by sampling from a Bernoulli distribution. This distribution takes

[7] The sampling procedure is nearly identical irrespective of whether the datum is held-out or observed, except that in the former, the likelihood is not calculated.

a single parameter, $0 \leq \beta \leq 1$, which corresponds to the probability of switching the parameter setting or keeping it the same:

(40) $\iota'_{xc} \mid \iota_{xc} \sim \textsc{Bernoulli}(\beta)$

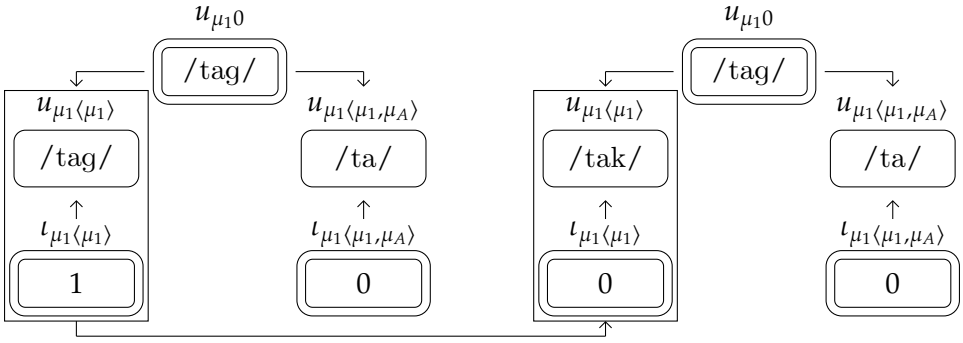
Suppose that the model samples to reverse the identity value. As the current identity value is set to $\iota_{\mu_1\langle\mu_1\rangle} = 1$, the newly-proposed identity value is set to $\iota'_{\mu_1\langle\mu_1\rangle} = 0$.[8]

Once the algorithm has sampled the identity value, it must now sample a new contextual UR for each lexeme in the lexical context. This is done by independently proposing a new contextual UR given the current contextual UR $Q(u'_{xc} \mid u_{xc})$.[9] The proposal distribution is based on the edit distance between the proposed and current contextual UR times the sensitivity parameter $0 \leq \phi$, as shown in (41).

(41) $u'_{xc} \mid u_{xc} \sim \textsc{Multinomial}(-\phi d_{u_{xc}})$

This multinomial distribution biases the algorithm to propose new contextual URs that are similar in form to the current contextual UR.[10] Let us say that the model happens to propose the new contextual UR $u_{\mu_1\langle\mu_1\rangle}$ = /tak/. This nets us the fully-specified proposed joint hypothesis in (42).

(42) Fully-specified proposed joint hypothesis for lexical context $\langle\mu_1\rangle$



Once a new hypothesis has been proposed, the sampling algorithm must decide whether to accept and update the model with the new hypothesis or reject it and keep the original hypothesis. To do this, the algorithm calculates the likelihood and prior. Taking the product of these values gives us the unnormalized WEIGHTED POSTERIOR of each hypothesis.

[8] The value of this parameter β does not directly affect the predictions of the model, but rather affects how good the sampled distribution estimates the true distribution. Moving forward, I fix $\beta = 0.5$.

[9] In order to speed up convergence, I restrict the sample space to only include contextual URs that achieve a prior higher than one, i.e. if $\iota_{xc} = 1$, the model sets the contextual UR equal to the prototype UR.

[10] As above, this parameter only affects the convergence of the algorithm. Moving forward, I fix $\phi = 3$.

In addition to the weighted posterior, the model calculates the TRANSITION PROBABILITIES Q of each hypothesis. The transition probabilities refer to the probabilities that we would have proposed one hypothesis given the other, and correspond directly to the proposal distributions described earlier. In this case, it would be the probability of proposing one identity hypothesis given the other, $Q(\iota'_{xc} \mid \iota_{xc})$ and $Q(\iota_{xc} \mid \iota'_{xc})$, as well as the probability of proposing some new contextual UR given the current contextual UR and vice versa, $Q(u'_{xc} \mid u_{xc})$ and $Q(u_{xc} \mid u'_{xc})$.

After computing the weighted posteriors and transition probabilities, the algorithm examines the ratio R of the product of both probabilities for each hypothesis, computed as in (43).

$$R = \frac{P(o_c \mid m, u'_c) \prod_{x \in x_c} P(u'_{xc} \mid u_{x0})\, P(\iota'_{xc})\, Q(\iota_{xc} \mid \iota'_{xc})\, Q(u_{xc} \mid u'_{xc})}{P(o_c \mid m, u_c) \prod_{x \in x_c} P(u_{xc} \mid u_{x0})\, P(\iota_{xc})\, Q(\iota'_{xc} \mid \iota_{xc})\, Q(u'_{xc} \mid u_{xc})} \tag{43}$$

Samples are then accepted or rejected via an acceptance function A, typified in (44), which returns a probability of acceptance based on this ratio.

$$A(h', h) = \begin{cases} 1 & \text{if } R > 1 \\ R & \text{otherwise} \end{cases} \tag{44}$$

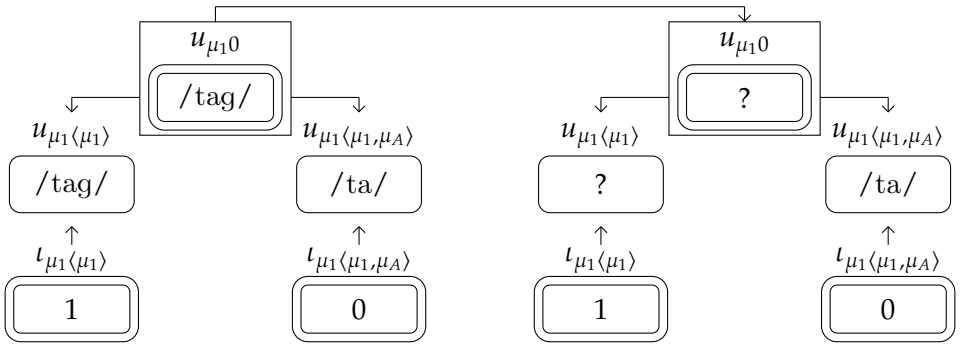
Newly sampled hypotheses are accepted categorically if it has a higher weighted posterior than the current UR. Otherwise, the new hypothesis is accepted as a function of the ratio of the weighted posterior and transition probabilities of the two hypotheses. Given that the model is biased against generating new contextual URs, the proposed hypothesis in our example has a lower weighted posterior. Let us suppose the model rejects the proposed hypothesis. As a result, the contextual UR $u_{\mu_1\langle\mu_1\rangle}$ remains unchanged. This continues for h iterations. Whichever hypothesis is recorded by the model by the end of those iterations is the parameter setting for that Gibbs iteration.

Estimating the conditional probability over prototype URs

Once all the contextual URs are sampled, the model then samples a prototype UR for each lexeme.[11] This process is schematized for the lexeme μ_1 in (45).

[11] The order in which parameters are updated do not matter. This order was selected primarily to illustrate how prototype UR hypotheses can indirectly affect the form of the contextual URs.

(45) Sampling procedure for the prototype UR $u_{\mu_1 0}$

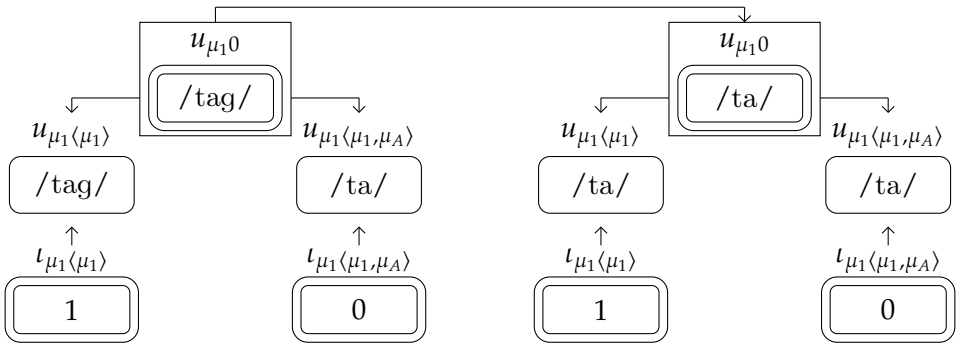


Suppose the model is are sampling a hypothesis for the prototype UR $u_{\mu_1 0}$. In each Metropolis-Hastings iteration, the sampling algorithm proposes a new hypothesis by sampling from a proposal distribution $Q(u'_{x0} \mid u_{x0})$. The distribution I adopt is identical to the one given in (41).

(46) $u'_{x0} \mid u_{x0} \sim \textsc{Multinomial}(-\phi d_{u_{x0}})$

Let us say that the model samples the prototype UR /ta/. Since the model denoted in the previous stage which contextual URs will use the default form, all of those contextual URs will also change in the newly proposed UR hypothesis. In the example given in (45) above, the lexical context $\langle\mu_1\rangle$ is marked as using the default prototype UR; as such, the associated contextual UR is also changed to /ta/. This can be seen in (47).

(47) Updated prototype UR hypothesis for μ_1



In order to decide whether to accept and update the hypothesis or to reject it and keep the original hypothesis, the algorithm once again calculates the weighted posterior and transition probability of each. Note that what goes into the calculation is slightly different from sampling a contextual UR; the algorithm calculates and compares the prior of the prototype UR $P(u_{x0})$ as well as the likelihood of the default contextual URs given the new

default form $P(u_{xc} \mid u_{x0})$. Contextual URs that do not use the default form generate the same likelihoods and thus cancel out. It then examine the ratio of the probabilities of each hypothesis, as computed in (48).

$$R = \frac{\prod_c P(o_c \mid m, u'_c) P(u'_{xc} \mid u'_{x0})\, P(u'_{x0})\, Q(u_{x0} \mid u'_{x0})}{\prod_c P(o_c \mid m, u_c) P(u_{xc} \mid u_{x0})\, P(u_{x0})\, Q(u'_{x0} \mid u_{x0})} \tag{48}$$

Lastly, the algorithm decides whether to accept or reject this sample via the same acceptance function laid out in (44) above. This process is repeated h times, from which the model then returns and updates the model with the final hypothesis at the end of the hth iteration.

2.3.3 Summary of the inference procedure

In the previous sections, I went over how the model performs inference over the lexicon and grammar by way of two sampling algorithms: Gibbs sampling and Metroplis-Hastings within Gibbs sampling. I summarize the entire inference process in (49).

(49)
a. Initialize the UR u and mapping m hypotheses
b. For each lexical context c, sample a new identity value ι_c and UR u_c
   (i) Sample a new identity parameter $\iota'_{xc} \sim \textsc{Bernoulli}(\beta)$
   (ii) Sample a new contextual UR $u'_{xc} \mid u_{xc} \sim \textsc{Multinomial}(-\phi d_{u_{xc}})$
   (iii) Calculate the ratio of the weighted posteriors and transition probabilities for each of the hypotheses and accept or reject the new hypothesis via the acceptance function
   (iv) Repeat (49-b-i-iii) for h iterations; return the last UR hypothesis
c. For each lexeme x, sample a prototype UR u_{x0}
   (i) Sample a new prototype UR $u'_{x0} \mid u_{x0} \sim \textsc{Multinomial}(-\phi d_{u_{x0}})$
   (ii) Calculate the ratio of the weighted posteriors and transition probabilities for each of the hypotheses and accept or reject the new hypothesis via the acceptance function
   (iii) Repeat (49-c-i-iii) for h iterations; return the last UR hypothesis
d. Sample a mapping hypothesis given the other parameters $P(m \mid u, \iota, o)$ and update the model with the sampled mapping hypothesis
e. Save the completely updated model
f. Repeat (49-a-e) for g iterations and return the sampled hypotheses
g. Remove the first half of the samples drawn, and return every bth sample

Table 2-2: Sample language demonstrating final-devoicing.

Alternating				Non-alternating			
tak	$\langle \mu_1 \rangle$	taga	$\langle \mu_1, \mu_A \rangle$	tak	$\langle \mu_5 \rangle$	taka	$\langle \mu_5, \mu_A \rangle$
tat	$\langle \mu_2 \rangle$	tada	$\langle \mu_2, \mu_A \rangle$	tat	$\langle \mu_6 \rangle$	tata	$\langle \mu_6, \mu_A \rangle$
???	$\langle \mu_3 \rangle$	kaga	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_7 \rangle$	kaka	$\langle \mu_7, \mu_A \rangle$
???	$\langle \mu_4 \rangle$	kada	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_8 \rangle$	kata	$\langle \mu_8, \mu_A \rangle$

2.4 Encoding the Lexicophonological Space and Predicting Forms

The inference process described above ends with an estimate of the joint posterior distribution over lexicons and grammars. How does this joint posterior relate to previous discussions regarding the grammatical space given in Chapter 1? How do we evaluate the model's performance given this space? To answer these questions, I will refer to the final-devoicing language one last time. The sample language is repeated once again from Table 2-2.

As a reminder, the ORs in the data were originally generated via the application of a single phonological process of final-devoicing: O → T / _#. Depending on the UR of the stem, we generate either the alternating or non-alternating paradigms: stems with an underlying final voiced consonant produce the alternating paradigms, while stems with an underlying final voiceless consonant produce the non-alternating paradigms. For example, the URs for the lexical contexts $\langle \mu_1, \mu_A \rangle$ and $\langle \mu_1 \rangle$ are /tag-a/ and /tag/, respectively. Applying final-devoicing to these URs produces the alternating ORs [taga] and [tak], respectively. In contrast, the URs for the lexical contexts $\langle \mu_5, \mu_A \rangle$ and $\langle \mu_5 \rangle$ are /tak-a/ and /tak/, respectively. Final devoicing applies vacuously to theses forms, resulting in the non-alternating ORs [taka] and [tak], respectively. This discussion is replicated in the following derivation table:

(50) Lexicon and grammar used to generate the final-devoicing ORs. From left to right, the lexical contexts for each UR are: $\langle \mu_1, \mu_A \rangle$, $\langle \mu_1 \rangle$, $\langle \mu_5, \mu_A \rangle$, $\langle \mu_5 \rangle$

Underlying Form	/tag-a/	/tag/	/tak-a/	/tak/
O → T / _#	–	[tak]	–	–
Observed Form	[taga]	[tak]	[taka]	[tak]

While the ORs of the data were generated via the lexicon and grammar specified in (50), what joint hypotheses are compatible with the data? For instance, another hypothesis

compatible with the data is one in which the alternations are encoded by the lexicon to the exclusion of any learned phonological processes. This is illustrated in (51) below.

(51) Alternative lexicon and grammar that can generated the final-devoicing ORs. From left to right, the lexical contexts for each UR are: $\langle \mu_1, \mu_A \rangle$, $\langle \mu_1 \rangle$, $\langle \mu_5, \mu_A \rangle$, $\langle \mu_5 \rangle$

Underlying Form	/tag-a/	/tak/	/tak-a/	/tak/
∅	–	–	–	–
Observed Form	[taga]	[tak]	[taka]	[tak]

As we can see, the alternation observed in the ORs is accounted for not by a phonological process, but rather through contextual allomorphy. Each of the discussed hypotheses in turn potentially make very different predictions on held-out lexical contexts. Let us walk through the predictions of each.

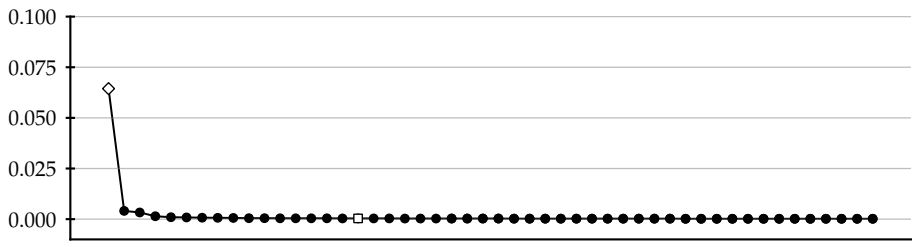
Suppose the model posits a rule hypothesis consisting only of final-devoicing, as in (50) above. Given the OR [kaga] for the lexical context $\langle \mu_3, \mu_A \rangle$, the model would be able to correctly determine that the UR for the lexical context is /kag-a/; otherwise, it would fail to capture the data. As the model possesses a prior biasing the model to reuse the same UR rather than to generate new context-specific ones, the hypothesized UR for the lexeme μ_3 extends to other contexts, including the held-out lexical context $\langle \mu_3 \rangle$. Applying final devoicing to the UR /tag/ results in the form [tak], consistent with the rest of the observed data. The same story applies to the non-alternating paradigms as well: the model would be able to determine that the UR for the lexical context $\langle \mu_7, \mu_A \rangle$ is /kak-a/, or it would fail to reconstruct the observed OR [kaka] in the data. The prior would once again extend the UR hypothesis beyond the observed context, allowing the model to posit the UR /kak/ for the held-out lexical context $\langle \mu_7 \rangle$. This is demonstrated in (52).

(52) Predictions on held-out data under the lexicon and grammar given in (50). From left to right, the lexical contexts for each UR are: $\langle \mu_3, \mu_A \rangle$, $\langle \mu_3 \rangle$, $\langle \mu_7, \mu_A \rangle$, and $\langle \mu_7 \rangle$

Underlying Form	/kag-a/	/kag/	/kak-a/	/kak/
O → T / _#	–	[kak]	–	–
Observed Form	[kaga]	[kak]	[kaka]	[kak]

Suppose instead that the model decides not to posit any phonological mappings, as in (51). If the model selects the contextual UR /kag/ for the lexical context $\langle \mu_7, \mu_A \rangle$, then the prior

Figure 2-7: Descending posterior probabilities of the top 50 hypotheses. The white diamond corresponds to the intended grammar in (50). The white square corresponds to the intervocalic voicing grammar in (54).



will bias the model to reuse that UR for the held-out lexical context $\langle \mu_7 \rangle$. This, however, produces the alternative, unintended output *[kag], as the generalization observed in the complete paradigm fails to extend to new environments. This is shown in (53).

(53) Predictions on held-out data under the lexicon and grammar given in (51). From left to right, the lexical contexts for each UR are: $\langle \mu_3, \mu_A \rangle$, $\langle \mu_3 \rangle$, $\langle \mu_7, \mu_A \rangle$, and $\langle \mu_7 \rangle$

Underlying Form	/kag-a/	/kag/	/kak-a/	/kak/
∅	–	–	–	–
Observed Form	[kaga]	*[kag]	[kaka]	[kak]

Which of the hypotheses does the model prefer? As it turns out, the original hypothesis in (50) has the higher posterior probability. To illustrate this, I plot the posterior disribution in order of greatest to lowest magnitude, given in Figure 2-7. I mark the point corresponding to the target grammar with a white diamond.

This observation emerges as the intended hypothesis maximizes both the likelihood and the prior: this lexicon-grammar pair perfectly captures the ORs of the data while also minimizing the number of unique contextual URs needed to generate the forms; in fact, the hypothesis only posits a single UR for each lexeme for all contexts it is found in. In contrast, the alternative grammar, while perfectly fitting the data it has been trained on, must do so at the cost of the prior; for each form that it must memorize, it must do so by postulating a new contextual UR for the lexical context. As a consequence, it has a comparably lower posterior – so low, in fact, that the hypothesis is not visible on the plot.

However, these are not the only hypotheses on the market; we can indeed see that while the intended hypothesis has the *highest* posterior probability, it occupies only a small portion of the entire space, having a posterior of 6.5%. This still leaves 93.5% of the distribution to allocate. Moreover, the posterior distribution is not restricted to only

hypotheses *consistent* with the data, but also hypotheses that are nearly consistent. For example, consider a hypothesis in which final-devoicing is not learned, but rather an alternative analysis of intervocalic voicing is posited instead. Under this analysis, the model predicts an intervocalic voiced consonant regardless of whether the original paradigm was an alternating or non-alternating paradigm. This is demonstrated in (54).

(54) Lexicon and grammar under an intervocalic voicing analysis. From left to right, the lexical contexts for each UR are: $\langle \mu_1, \mu_A \rangle$, $\langle \mu_1 \rangle$, $\langle \mu_5, \mu_A \rangle$, $\langle \mu_5 \rangle$

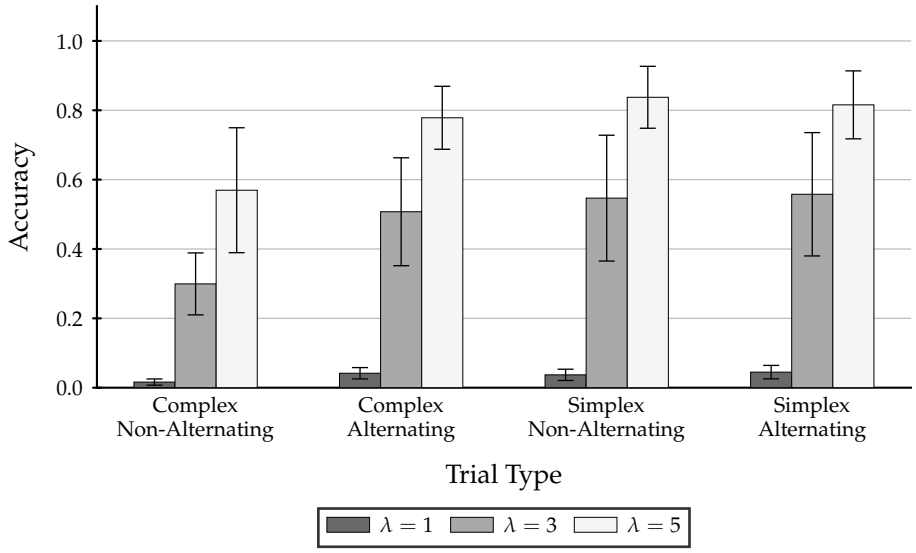
Underlying Form	/tak-a/	/tak/	/tak-a/	/tak/
O → D / V_V	[taga]	–	[taga]	–
Observed Form	[taga]	[tak]	*[taga]	[tak]

While the hypothesis gets the some of the ORs of the data wrong, the level of deviation is quite small; due to the fact that the voiceless and voiced segments are very similar to one another, the cost of incorrectly producing the incorrect consonant is not too high as to exclude the hypothesis. Moreover, despite the joint lexicon-grammar hypothesis having a relatively lower likelihood, it continues to maximize the prior; the lexicon still posits only a single UR for all the contextual URs each lexeme is found in. It consequently has a higher posterior than expected, which I plot as a white square in Figure 2-7.

From the discussion above, we can take note of two major observations. ① Different hypotheses in the hypothesis space have different relative posteriors based on the interplay of the likelihood, which determines the fit of the ORs relative to that hypothesis' ERs, and the prior, which determines how probable the hypothesis is before observing the data. ② Each hypothesis makes different predictions on both the observed data, but more importantly, on held-out forms, the latter of which is directly evaluated in artificial grammar learning experiments. The posterior distribution thus encodes the original notion of the grammatical space by relating the data to the entire space of possible lexicon-grammar hypotheses. However, unlike the typical notion of grammatical spaces, which examines only the space of consistent grammars under a uniform prior, the noisy channel and generative model over the lexicon allows the model to both consider nearly consistent lexicons in addition to grammars, and shifts the density of the posterior to each hypothesis differently. As such, the posterior behaves more in terms of a weighted inferred lexicophonological space, where each hypothesis has slightly different weight when making a prediction.

In order to make predictions on held-out data given the inferred space of hypotheses, I calculate the weighted average expected output for an unobserved lexical context given this space, typified in (55). This is known as the posterior predictive distribution.

Figure 2-8: Posterior predictive for the final-devoicing language. Top: $\lambda = 5$. Bottom: $\lambda = 1$.



(55) $P(e_c^* \mid o) = \sum_{m,u} \mathbb{1}[e_c^* = m(u_c)]\, P(m, u \mid o)$

Note that there is a distinction between the inferential model presented in the previous section and the the predictive model presented here. During inference, the model assumes that the ORs were generated by way of a noisy channel $P(o \mid e)$; however, when generating a prediction, the model does not apply the noisy channel, instead utilizing the ER $P(e \mid m, u)$ directly, in the absence of noise. The noisy channel component is intended to be reflective of the learners' beliefs regarding the nature of the data they are seeing. In other words, the noisy channel only indirectly affects the joint posterior distribution and does not directly affect how the learner in turn produces a form.

Let us examine what the model predicts for the held-out forms for this language given the posterior distribution inferred. This is shown in Figure 2-8. I will compare the model using three different parameter settings for λ in order of decreasing noise: $\lambda \in \{1, 3, 5\}$. In order to evaluate the predictive distribution for each hyperparameter setting, I report the expected predictive probability assigned to each of four context types: ① observed complex non-alternating lexical contexts, e.g. $\langle \mu_7, \mu_A \rangle$, ② observed complex alternating lexical contexts, e.g. $\langle \mu_3, \mu_A \rangle$, ③ unobserved simplex non-alternating lexical contexts, $\langle \mu_7 \rangle$, and ④ unobserved simplex alternating lexical contexts, $\langle \mu_3 \rangle$. The expected probability is defined as the probability the model assigns to the form expected under the intended hypothesis. Here, this corresponds to a language in which segments contrast in voicing intervocallically, e.g. [kaga] and [kaka] for the lexical contexts $\langle \mu_3, \mu_A \rangle$ and $\langle \mu_7, \mu_A \rangle$, but not

word-finally, e.g. [kat] for the lexical contexts $\langle \mu_3 \rangle$ and $\langle \mu_7 \rangle$.

Increasing the value of λ decreases the level of deviation from the observed surface data. In other words, increasing λ prioritizes the likelihood with respect to the prior when computing the weighted posterior. We can observe the direct consequence of this in the likelihood of the observed forms: as the model allows for less noise, the probability assigned to the expected forms increases. In other words, there is a balancing act between the prior, the likelihood, and the entire lexicophonological space.

2.5 Conclusion

In this chapter, I proposed and formalized the noisy-channel lexicophonological learner, explicating how the model generates hypotheses and performs inference over its parameters. I discussed how this model encodes a refined version of the grammatical space via the posterior distribution and posterior predictive distribution: the model calculates the posterior, which assigns a probability to every hypothesis in the space. This weighted inferred space is then used to predict held-out lexical contexts, which we then use to evaluate model performance, and thereby learnability. In the next chapter, I will apply this model to two recent artificial grammar learning experiments that examined the learnability of process interactions and advance an alternative explanation to the observed learnability asymmetry: certain phenomena are easier to learn as a consequence of how many grammars are consistent with or nearly consistent with the data produced by each interaction.

Chapter 3

How the Lexicophonological Space Shapes Generalization

In this chapter, I compare the predictions of the noisy-channel model to the results of two artificial language learning experiments. Despite each experiment reaching seemingly incompatible conclusions, I demonstrate that the model successfully captures the qualitative asymmetries observed in both under a unified account: differences in the generalizability of a pattern emerge due to differences in the number of consistent or nearly consistent lexicon-grammar hypotheses associated with each pattern. The space of consistent and nearly consistent hypotheses is shown to be influenced in large part by the relative frequencies of ORs containing each alternant in the distribution, thus supporting the hypothesis that the variation in learnability emerges as a consequence of the distribution of ORs given to the learner as opposed to formal properties of the process interaction itself. The chapter is organized into the following three sections. ① I discuss how the model will be parameterized throughout the remainder of the chapter. ② I summarize and computationally model a recent artificial language learning experiment by Prickett (2019) who investigated the learnability of four process interactions: that of a bleeding, feeding, counter-bleeding, and counter-feeding interaction. ③ I introduce and computationally model a different artificial language learning experiment by Kim (2012) that involves the same underlying feeding and counter-feeding interactions observed in the previous experiment but produced seemingly contradictory results. I argue that the apparent discrepancy emerges as a consequence of differences in the distributions of the ORs of each artificial language which results in a shift in the inferred lexicophonological space and thus the predictions given that space. The two experimental sections are further distributed into three subsections. Ⓐ I go over how the ORs of each artificial language are generated and organized. Ⓑ I go over the hypothesis

Table 3-1: Parameterization of the noisy-channel lexicophonological model.

Parameter	Setting	Effect on the Model
θ	0.5	Prefer shorter prototype URs.
α	0.75	Prefer to reuse the prototype UR.
ψ	5	Prefer to generate contextual URs similar to the prototype UR.
λ	*varies*	Adjusts how similar the ERs must be to their ORs.

space of URs and rules used by the learner to model the experiment. © I compare the results of the model to the results of each respective experiment.

3.1 Model Parameterization

The lexicon and noisy channel are associated with several parameters that affect the model's behavior. For example, the lexicon is parameterized by the terms θ, α, and ψ, the first of which dictates the length and identity of the prototype URs, and the latter two of which dictate how faithful the contextual URs will be to the default prototype UR. The noisy channel in turn is parameterized by the term λ, which adjusts the level of noise in the model. In order to examine the relationship between the distribution of the data and the noisy channel, I fix all the parameters of the model except for λ. A summary of each parameter and what the parameter setting will be is given in Table 3-1.

3.2 Modeling Prickett (2019)

Prickett sought to investigate the learnability of different process interactions by performing an artificial grammar learning experiment. The experiment involved training 48 participants in one of four toy languages: a bleeding language, a counter-bleeding language, a feeding language, and a counter-feeding language.

Each language is composed of several morpho-phonological paradigms consisting of the stem in isolation as well as three conjugated forms: ① the stem with an /-i/ suffix, ② the stem with an /-a/ suffix, and ③ the stem with both the /-i/ and /-a/ suffixes in some order. All of the languages consist of two paradigm types: alternating coronal-final stems, e.g. /imat/ and /imad/, and non-alternating velar-final stems, e.g. /imak/ and /imag/. To simplify discourse, I limit the discussion over coronal-final stems to [t] and velar-final stems to [k]. Each OR in a paradigm is generated by sequentially applying in some order the deletion and palatalization processes outlined in (1) to each UR.

(1) a. DELETION: V → ∅ / _V
b. PALATALIZATION: t → tʃ / _i

Each language differs only in the order in which the /-a/ and /-i/ suffixes attach to the stem as well as the order in which the processes apply to the concatenated URs. Having /-a/ attach before /-i/ generates the feeding languages, whereas having /-i/ attach before /-a/ generates the bleeding languages. Applying deletion followed by palatalization results in the transparent feeding and bleeding languages, whereas applying palatalization followed by deletion results in the opaque counter-bleeding and counter-feeding languages. Note that these languages are identical to the revised Baković languages introduced in Chapter 1. Sample paradigms illustrating each of the four languages are given in Table 3-2. The table is organized into a 2×2 grid corresponding to each language. Each language is further divided into two sub-tables, corresponding to the alternating coronal-final stem paradigms and the non-alternating velar-final stem paradigms. Languages within a column possess the same order of suffixation, but opposite rule ordering. Languages within a row possess the same rule ordering, but opposite order of suffixation.

Prickett organized the lexical contexts of each language into distinct TRIAL TYPES based on which phonological process or processes are expected to apply to its respective UR. The FAITHFUL FORMS are associated with lexical contexts in which no phonological transformation is expected to occur. These correspond to lexical contexts in which a coronal-final stem combines with the /-a/ suffix, such as /imat-a/, or a velar-final stem combines with either the /-a/ or /-i/ suffix, such as /imak-a/ and /imak-i/. The URs /imat-a/ or /imak-i/ do not satisfy the structural description for either process to apply, and thus are expected to surface faithfully. The PALATALIZING FORMS correspond to lexical contexts in which only the palatalization process is expected to apply, such as those in which a coronal-final stem combines with the /-i/ suffix, such as /imat-i/. The UR /imat-i/ satisfies the requisite structural description for palatalization, and thus is expected to surface with the respective OR containing the palatalized segment, i.e. [imatʃi]. The DELETING FORMS refer to lexical contexts in which a velar-final stem combines with both suffixes /-a/ and /-i/, such as /imak-a-i/ or /imak-i-a/. The URs /imak-a-i/ and /imak-i-a/ both satisfy the conditions for vowel deletion, and thus are expected to surface with the target vowel deleted, i.e. [imaki] and [imaka]. Finally, the INTERACTING FORMS correspond to lexical contexts in which the two individual processes interact under the original grammar, such as lexical contexts containing coronal-final stems combined with both the /-i/ and /-a/ suffixes in some order, e.g. /imat-a-i/ or /imat-i-a/. The OR varies depending on the language, as enumerated above.

Table 3-2: Sample URs and ORs for each of the revised Baković (2011) languages in Prickett (2019).

	Bleeding Language				Feeding Language			
Underlying Form	/t/	/t-a/	/t-i/	/t-i-a/	/t/	/t-a/	/t-i/	/t-a-i/
V → ∅ / _V	–	–	–	[ta]	–	–	–	[ti]
t → tʃ / _i	–	–	[tʃi]	–	–	–	[tʃi]	[tʃi]
Observed Form	[t]	[ta]	[tʃi]	[ta]	[t]	[ta]	[tʃi]	[tʃi]
Underlying Form	/k/	/k-a/	/k-i/	/k-i-a/	/k/	/k-a/	/k-i/	/k-a-i/
V → ∅ / _V	–	–	–	[ka]	–	–	–	[ki]
t → tʃ / _i	–	–	–	–	–	–	–	–
Observed Form	[k]	[ka]	[ki]	[ka]	[k]	[ka]	[ki]	[ki]

	Counter-bleeding Language				Counter-feeding Language			
Underlying Form	/t/	/t-a/	/t-i/	/t-i-a/	/t/	/t-a/	/t-i/	/t-a-i/
t → tʃ / _i	–	–	[tʃi]	[tʃia]	–	–	[tʃi]	–
V → ∅ / _V	–	–	–	[tʃa]	–	–	–	[ti]
Observed Form	[t]	[ta]	[tʃi]	[tʃa]	[t]	[ta]	[tʃi]	[ti]
Underlying Form	/k/	/k-a/	/k-i/	/k-i-a/	/k/	/k-a/	/k-i/	/k-a-i/
t → tʃ / _i	–	–	–	–	–	–	–	–
V → ∅ / _V	–	–	–	[ka]	–	–	–	[ki]
Observed Form	[k]	[ka]	[ki]	[ka]	[k]	[ka]	[ki]	[ki]

In order to evaluate how well participants were doing in learning the phonology of the language, Prickett performed a forced-choice task. In each trial, participants were given a stem in isolation paired with an image representing its meaning. The participant was then given an image representing a complex lexical context and were then asked to select between one of two options. One of the options corresponds to the intended response, or the OR consistent with the original phonological grammar. For example, the intended response for the palatalizing trial, e.g. /imat-i/, would be the form in which the palatalized consonant surfaces, e.g. [imatʃi], while the intended response for the deleting trial, e.g. /imak-a-i/ or /imak-i-a/, would be the form in which the first vowel deletes, e.g. [imaki] and [imaka]. The other option corresponds to the alternate response, or the form expected if the learner either underlearned or mislearned the language. The alternate responses for the faithful trials consist of forms containing a palatalized consonant; for example,

Table 3-3: Schematized trial types and choices for each of the revised Baković (2011) languages in in Prickett (2019). The rows in each sub-table corresponds to the trial type, expected UR, intended response, and alternate response.

Bleeding Language

/t/	Faithful /t-a/	Palatal. /t-i/	Interacting /t-i-a/
	[ta]	[tʃi]	[ta]
	*[tʃa]	*[ti]	*[tʃa]

/k/	Faithful /k-a/	Faithful /k-i/	Deleting /k-i-a/
	[ka]	[ki]	[ka]
	*[tʃa]	*[tʃi]	*[kia]

Feeding Language

/t/	Faithful /t-a/	Palatal. /t-i/	Interacting /t-a-i/
	[ta]	[tʃi]	[tʃi]
	*[tʃa]	*[ti]	*[ti]

/k/	Faithful /k-a/	Faithful /k-i/	Interacting /k-a-i/
	[ka]	[ki]	[ki]
	*[tʃa]	*[tʃi]	*[kai]

Counter-bleeding Language

/t/	Faithful /t-a/	Palatal. /t-i/	Interacting /t-i-a/
	[ta]	[tʃi]	[tʃa]
	*[tʃa]	*[ti]	*[ta]

/k/	Faithful /k-a/	Faithful /k-i/	Deleting /k-i-a/
	[ka]	[ki]	[ka]
	*[tʃa]	*[tʃi]	*[kia]

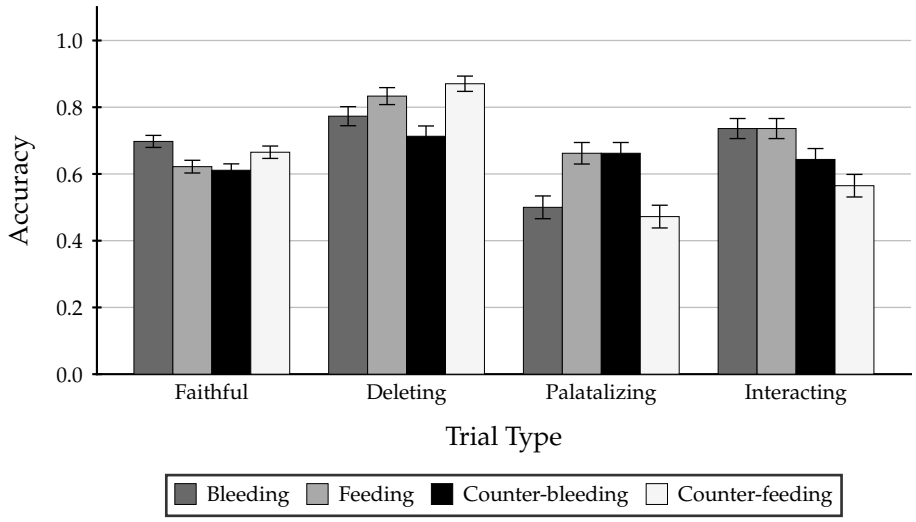
Counter-feeding Language

/t/	Faithful /t-a/	Palatal. /t-i/	Interacting /t-a-i/
	[ta]	[tʃi]	[ti]
	*[tʃa]	*[ti]	*[tʃi]

/k/	Faithful /k-a/	Faithful /k-i/	Interacting /k-a-i/
	[ka]	[ki]	[ki]
	*[tʃa]	*[tʃi]	*[kai]

the alternate responses for the forms /imat-a/ and /imak-i/ are *[imatʃa] and *[imatʃi], respectively. The alternate responses for the palatalizing trials consist of forms containing the unpalatalized consonant; for example, the alternate response for the form /imat-i/ is *[imati]. The alternate responses for the deleting trials consist of forms containing both vowels; for example, the alternate responses for the forms /imak-a-i/ and /imak-i-a/ are *[imakai] and *[imakia], respectively. Lastly, the alternate responses for the interacting trials consist of forms in which the opposite rule application applied; for example, the alternate responses for the interacting trials for the feeding and bleeding languages are *[imati] and *[imatʃa], identical to the intended output of the interacting trial for the counter-feeding and counter-bleeding languages, respectively. The trial types and associated schematized choices for each are summarized in Table 3-3.

Figure 3-1: Average aggregate accuracy for each trial type by language in Prickett (2019). Error bars correspond to 95% confidence intervals computed over all data points per condition.



The experiment was distributed into two phases: the training and testing phase. In the training phase, participants were given 36 complete paradigms, and feedback was given as to whether or not the response given is intended under the original lexicon and grammar. In the testing phase, participants were presented 36 incomplete novel paradigms and feedback was not given. The stems for each paradigm correspond to each of the permutations generated of the form {i, u}{m, n, l}{i, a, u}{t, k, d, g}. Performance was evaluated based on the relative frequency participants chose the intended response versus the alternate response on the novel incomplete paradigms.

3.2.1 Experimental results

The results of the experiment are shown in Figure 3-1. Statistically, Prickett uncovered two significant results. First, in the palatalizing trials, Prickett found that participants trained on the feeding and counter-bleeding languages had significantly higher accuracy than those trained on the bleeding and counter-feeding languages, corresponding to the asymmetry predicted under the Maximal Utilization bias. Second, for the interacting trials, Prickett found that participants performed better when generating the bleeding and feeding versus the counter-bleeding and counter-feeding forms, mimicking the asymmetry predicted under the Transparency bias.

The results seem to indicate two important observations. First, certain interactions appear to be easier to replicate and generalize than others across all trial types; partici-

pants trained on the feeding interaction performed either not significantly differently, or significantly better, than participants trained on the counter-feeding interaction. Second, evaluating performance for a given process interaction is determined not only on how participants reliably extend generalizations to the interacting form, but also the forms applying each individual process. We see, for example, that participants performed well in the interacting trial for the bleeding language compared to those trained in the counter-bleeding language, but the asymmetry is reversed in the palatalizing trial.

Prickett compared the results of the experiment to predictions made under two different computational models: ① an Expectation Driven Learner of Harmonic Serialism augumented with Serial Markedness Reduction constraints (McCarthy 2000; Jarosz 2014, 2015, 2016), and a Sequence-to-Sequence neural network (Sutskever, Vinyals, & Le 2014; Kirov 2017; Kirov & Cotterell 2018). He showed that both models succeeded in replicating the empirical asymmetries observed in the experiment. Prickett analyzed the results as occurring due to MISLEARNING as a result of GRAMMATICAL AMBIGUITY: certain distributions may be ambiguous between multiple grammatical hypotheses, which can only be distinguished by a small set of forms in the data. For example, the bleeding language is ambiguous between a hypothesis in which deletion follows palatalization and a hypothesis in which only deletion is learned, with the only ORs that provides the crucial disambiguating evidence being the palatalizing forms. In contrast, the feeding language is less ambiguous as to what the correct grammatical hypothesis would be, providing evidence of the palatalization process in both the palatalizing forms as well as the interacting forms. More ambiguous data cause both models to take longer to acquire the correct generalization, resulting in mislearning via incomplete learning.

The computational models Prickett utilized made the simplifying assumption that the UR-OR pairs were given for free. However, while the ORs of the language were generated by applying a particular sequence of phonogical processes to a particular set of URs, this exact hypothesis is not the only manner in which the data could have been produced. The language could have also be generated through the application of different phonological processes, through memorization, or a combination of the two. For example, the interacting OR for the counter-feeding language [imati] could have been generated by having its UR be /imat-_-i/ with neither palatalization nor deletion applying. This form is derived not via the application of phonological processes, but rather through rote memorization, explicitly encoding the output as its UR. Alternatively, the interacting OR could have been generated from an input /imatʃ-_-i/ with a general depalatalization process that transforms [tʃ] to [t] across all contexts. Under this hypothesis, an output is formed via the application of a phonological process that is not a part of the original grammar used to generate the

Table 3-4: Inventory used to model the revised Baković (2011) languages in Prickett (2019)

	[cons]	[cor]	[dist]	[high]
t	1	1	-1	0
tʃ	1	1	1	0
k	1	-1	0	1
a	-1	-1	0	-1
i	-1	-1	0	1

data. As the data is compatible with a number of different hypotheses, many of which are completely disjoint from the grammar assumed to be learned, it is hard to interpret what was actually learned by participants.

In the following sections, I demonstrate that the noisy-channel model provides a formal account of why certain some languages in the experiment have higher accuracy in certain forms than others: different languages have a different number of consistent or nearly consistent hypotheses associated with them. We will see that this claim is compatible with the mislearning hypothesis, but is achieved in a slightly different way.

3.2.2 Computational overview

In this section, I explore what a model that jointly infers a lexicon and a grammar predicts given each of the four toy languages. Explicitly, I seek to investigate whether the model is able to produce the same qualitative empirical predictions observed in the experiment. In order to evaluate the performance of the model on this experiment, I must provide the model with the hypothesis space over the lexicon, the hypothesis space over the rules, and the set of data. I discuss each in turn in the following sections.

Hypothesis space over the lexicon

The segment and feature inventory used by the model is given in Table 3-4. This inventory is composed of 5 segments and 4 features.[1] The prototype and contextual URs can consist of any possible permutation of these segments.[2] Thus, the model can decide to encode the palatalized consonant as an underlying phoneme of the language /tʃ/, as an allophone of the phoneme /t/, or somewhere in between.

[1] Due to convergence reasons, the provided inventory lacks several segments used in the original experiment. Given the space of rules that I will use, the results and analysis will not be affected by their absence.

[2] As discussed in the previous chapter, I truncate the distribution to only generate prototype and contextual URs up to a certain upper bound. As before, I set the upper bound to be 5: $\textsc{max}(|u_{x0}|) = \textsc{max}(|u_{xc}|) = 5$.

Hypothesis space over the rules

I lay out the space of possible rules in (2). The rule space consists of a vowel deletion process as well as two generalized and contextual palatalization and depalatalization phenomena. This rule space is intended to both allow the model to posit potential alternative analyses of the data, such as one based on de-palatalization, as well to to partially mimic the behavior of OT-based analyses, by allowing the model to propose grammars that impose restrictions on surface contrasts.

(2) a. Vowel Deletion: V → ∅ / _V
b. Palatalization: t → tʃ / _i
c. Generalized Palatalization: t → tʃ
d. Depalatalization: tʃ → t / _{C, #}
e. Generalized Depalatalization: tʃ → t

Distribution of the data

The model was provided with the schematized dataset in Table 3-5. The data consists of both complete and incomplete paradigms of the coronal-final and velar-final stems.[3] The complete paradigms provides with model with examples of each individual phonological process as well as their interaction, mirroring the training phase of the experiment. The incomplete paradigms are used to assess the model's ability to generalize, and consist only of the stem in isolation. Given the stem, the model makes predictions on the held-out forms of the paradigm, mimicking the testing phase of the experiment.

Participants were exposed to 36 complete and 36 incomplete paradigms in the experiment; however, due to convergence reasons, the total amount of data given to the model must be truncated. I thus also ask the following question: does the absolute type frequency change the results of the model even if we keep the relative type frequency the same? To test this, I evaluate the model under two different distributions: the single paradigm distribution and the multiple paradigm distribution. Under the single paradigm distribution, the model is presented one complete and one incomplete paradigm for each paradigm type. This corresponds exactly to the dataset outlined in Table 3-5. Under the multiple paradigm distribution, the model is presented three complete and three incomplete paradigms of the same type. These distributions are summarized in Table 3-6.

In order to evaluate whether the model is capturing the same empirical asymmetries as

[3] Due to performance reasons, I opt to present the model only the voiceless t-final and k-final stems, to the omission of the voiced d-final and g-final stems. Their absence will not cause any differences in the observed results and interpretation of those results.

Table 3-5: Schematic dataset of the revised Baković (2011) language given to the model.

Bleeding Language

ikit	$\langle \mu_1 \rangle$	ikita	$\langle \mu_1, \mu_A \rangle$	ikitʃi	$\langle \mu_1, \mu_B \rangle$	ikita	$\langle \mu_1, \mu_B, \mu_A \rangle$
ikik	$\langle \mu_2 \rangle$	ikika	$\langle \mu_2, \mu_A \rangle$	ikiki	$\langle \mu_2, \mu_B \rangle$	ikika	$\langle \mu_2, \mu_B, \mu_A \rangle$
akit	$\langle \mu_3 \rangle$	???	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_3, \mu_B \rangle$	???	$\langle \mu_3, \mu_B, \mu_A \rangle$
akik	$\langle \mu_4 \rangle$	???	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_4, \mu_B \rangle$	???	$\langle \mu_4, \mu_B, \mu_A \rangle$

Feeding Language

ikit	$\langle \mu_1 \rangle$	ikita	$\langle \mu_1, \mu_A \rangle$	ikitʃi	$\langle \mu_1, \mu_B \rangle$	ikitʃi	$\langle \mu_1, \mu_A, \mu_B \rangle$
ikik	$\langle \mu_2 \rangle$	ikika	$\langle \mu_2, \mu_A \rangle$	ikiki	$\langle \mu_2, \mu_B \rangle$	ikiki	$\langle \mu_2, \mu_A, \mu_B \rangle$
akit	$\langle \mu_3 \rangle$	???	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_3, \mu_B \rangle$	???	$\langle \mu_3, \mu_A, \mu_B \rangle$
akik	$\langle \mu_4 \rangle$	???	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_4, \mu_B \rangle$	???	$\langle \mu_4, \mu_A, \mu_B \rangle$

Counter-bleeding Language

ikit	$\langle \mu_1 \rangle$	ikita	$\langle \mu_1, \mu_A \rangle$	ikitʃi	$\langle \mu_1, \mu_B \rangle$	ikitʃa	$\langle \mu_1, \mu_B, \mu_A \rangle$
ikik	$\langle \mu_2 \rangle$	ikika	$\langle \mu_2, \mu_A \rangle$	ikiki	$\langle \mu_2, \mu_B \rangle$	ikika	$\langle \mu_2, \mu_B, \mu_A \rangle$
akit	$\langle \mu_3 \rangle$	???	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_3, \mu_B \rangle$	???	$\langle \mu_3, \mu_B, \mu_A \rangle$
akik	$\langle \mu_4 \rangle$	???	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_4, \mu_B \rangle$	???	$\langle \mu_4, \mu_B, \mu_A \rangle$

Counter-feeding Language

ikit	$\langle \mu_1 \rangle$	ikita	$\langle \mu_1, \mu_A \rangle$	ikitʃi	$\langle \mu_1, \mu_B \rangle$	ikiti	$\langle \mu_1, \mu_A, \mu_B \rangle$
ikik	$\langle \mu_2 \rangle$	ikika	$\langle \mu_2, \mu_A \rangle$	ikiki	$\langle \mu_2, \mu_B \rangle$	ikiki	$\langle \mu_2, \mu_A, \mu_B \rangle$
akit	$\langle \mu_3 \rangle$	???	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_3, \mu_B \rangle$	???	$\langle \mu_3, \mu_A, \mu_B \rangle$
akik	$\langle \mu_4 \rangle$	???	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_4, \mu_B \rangle$	???	$\langle \mu_4, \mu_A, \mu_B \rangle$

those observed in the experiment, I focus on the model's capacity to produce held-out data. The experiment is a forced-choice task, but the noisy-channel model produces a posterior predictive that includes many ERs beyond the two given options. As such, I compute accuracy as in Prickett (2019), where the expected probabilities $P(e^* \mid o)$ are renormalized to only include each option of the forced-choice task:

$$\text{(3)} \qquad \textsc{Accuracy} = \frac{P(e^*_{\textsc{intended}} \mid o)}{P(e^*_{\textsc{intended}} \mid o) + P(e^*_{\textsc{alternate}} \mid o)}$$

3.2.3 Computational results

I present the results of the model for each type distribution in this section. This section is organized into two parts, corresponding to the single and multiple paradigm distributions

Table 3-6: Type distributions for modelling the Prickett (2019) experiment.

	Single Paradigm		Multiple Paradigm	
	Observed	Held-out	Observed	Held-out
Faithful	1	1	3	3
Deleting	1	1	3	3
Palatalizing	1	1	3	3
Interacting	1	1	3	3

discussed in Table 3-6. The results of the model for each distribution is discussed in turn in the following subsections below.

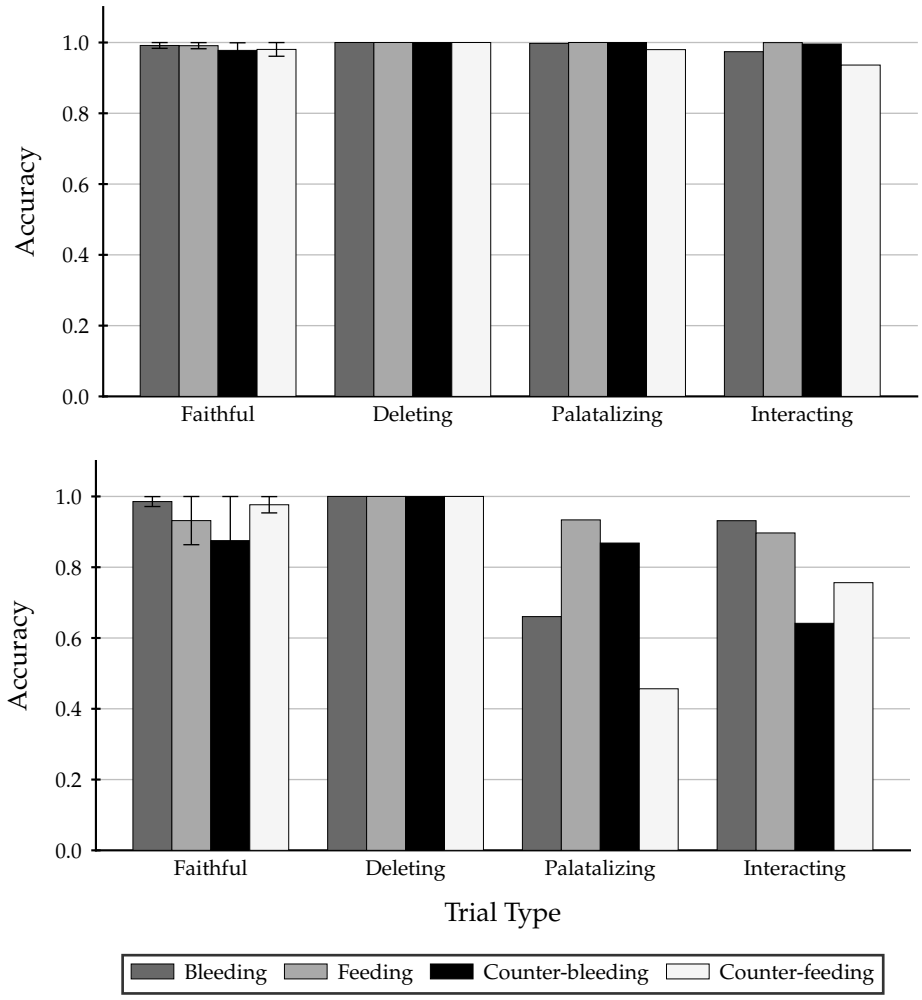
Single paradigm distribution results

I tested the model under two parameter settings over the noisy channel: one with ① little noise $\lambda = 10$ and ② moderate noise $\lambda = 3$. The results of the model on held-out lexical contexts for each trial type are given in Figure 3-2. We see that while the model fails to capture the empirical asymmetries with $\lambda = 10$, we indeed find that the noisy model with $\lambda = 3$ succeeds in capturing the main empirical observations made by Prickett.

I argue that the results emerge as a consequence of the number of lexicon and grammar pairs that are consistent or nearly consistent with the observed data; patterns generated by certain process interactions have more compatible hypotheses than others, and thus are easier to reproduce, even if the interaction itself is not learned. In order to better illustrate these effects, I will discuss first how the model under a low-noise parameterization *fails* to get the results before moving on to the model with a high-noise parameterization.

The model parameterized under low noise is intended to illustrate what predictions the model would make if it only considered hypotheses that perfectly generated the ORs given in the data. We see under this model three key observations. ① The model performs at near-ceiling across all four languages. This seems to suggest that the model indeed is capable of learning and generalizing all four patterns, given low enough noise. ② The model roughly achieves the Maximum Utilization bias observed in the experiment for the palatalizing trials, with the model performing (slightly) better on this trial when trained on the feeding and counter-bleeding interactions as opposed to the counter-feeding and bleeding interactions. However, ③ the model fails to get the expected Transparency bias in the interacting trials, instead getting the same preference observed in the palatalizing trials, with the maximally utilized interactions outperforming the non-maximally utilized interactions. Note that while the qualitative asymmetry expected under the Transparency

Figure 3-2: Average accuracy of the model for the Prickett experiment for each trial type by language under the SINGLE PARADIGM distribution. TOP: $\lambda = 10$. BOTTOM: $\lambda = 3$.



bias does not emerge here, certain properties expected under this bias do emerge; for example, note that the difference in absolute accuracy between the feeding and counter-feeding languages is larger than the difference observed between the bleeding and counter-bleeding languages. I will argue that the source of the observed asymmetry emerges due to the difference in the number of consistent lexicon and grammar pairs. This space of consistent hypotheses is largely determined by the INFORMATIVITY of each form, which can be formalized by the input and output provision and removal properties of each interaction, as first proposed by Baković and Blumenfeld (2022) and discussed in Chapter 1. I repeat the definitions of each in (4) and (5).

(4) Input Interactions: given two phonological processes A and B:

a. A input-provides B if there exists from mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $x \xrightarrow{B}$ is vacuous
   (iii) $y \xrightarrow{B}$ is not vacuous
   A creates a form in which B can apply where before, it could not.

b. A input-removes B if there exists some mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $x \xrightarrow{B}$ is not vacuous
   (iii) $y \xrightarrow{B}$ is vacuous
   A creates a form in which B cannot apply where before, it could.

(5) Output Interactions: given two phonological processes A and B:

a. A output-provides B if there exists some mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $\xrightarrow{B} x$ is vacuous
   (iii) $\xrightarrow{B} y$ is not vacuous
   There is a form y that A can create that B can also create without applying A.

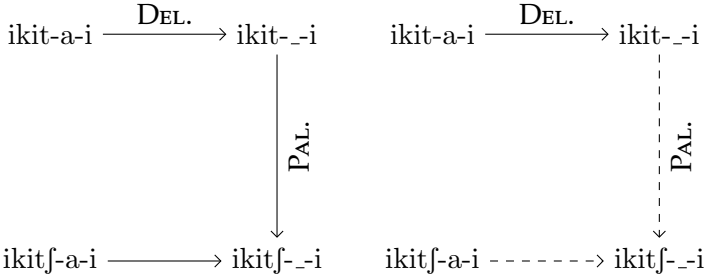
b. A output-removes B if there exists some mapping $x \xrightarrow{A} y$ such that:
   (i) $x \xrightarrow{A}$ is not vacuous
   (ii) $\xrightarrow{B} x$ is not vacuous
   (iii) $\xrightarrow{B} y$ is vacuous
   There is a form y that B can create that A cannot create without first applying B.

I will first go over the asymmetries observed between the feeding and counter-feeding languages before moving on to the bleeding and counter-bleeding languages.

Empirically, across the palatalizing and interacting trials, both participants in the experiment and the computational model performed better when trained on the feeding language than on the counter-feeding language. To examine whether there is a difference in the space of compatible lexicon-grammar pairs, I again enumerate the space of consistent hypotheses for the interacting ORs of each language from Chapter 1 in (6).[4]

[4] I omit the generalized palatalization and depalatalization processes in our discussion here. The reason is due to the fact that the generalized process only produces perfectly consistent results under very specific scenarios, i.e. when the generalized depalatalization process is followed at some point by the contextual palatalization process. They thus do not make a noticeable impact on the results observed here. I will revisit these rules when discussing the nearly consistent grammars later in the section.

(6) Space of compatible joint lexicons and grammars for the feeding OR [ikitʃi] (LEFT) and counter-feeding OR [ikiti] (RIGHT)



The reason for this lies in the relationship between the lexicophonological space and a form's informativity. The basic notion of informativity denotes that, the more informative a form is, the fewer lexicons and grammars compatible with that form. In other words, more informative forms restrict the space of compatible hypotheses more than uninformative forms. We can translate this property in terms of input and output provision and removal.

In order to illustrate the basic principle, as in Chapter 1, I provide sample form pairs that demonstrate the relevant properties of each input and output interaction for the feeding and counter-feeding ORs in (7) and (8),

(7) a. DELETION input-provides PALATALIZATION: $\langle x, y \rangle = \langle$ikitai, ikiti$\rangle$
   (i) $x \xrightarrow{\text{DEL.}}$ is not vacuous: /ikitai/ $\xrightarrow{\text{DEL.}}$ [ikiti]
   (ii) $x \xrightarrow{\text{PAL.}}$ is vacuous: /ikitai/ $\xrightarrow{\text{PAL.}}$ [ikitai]
   (iii) $y \xrightarrow{\text{PAL.}}$ is not vacuous: /ikiti/ $\xrightarrow{\text{PAL.}}$ [ikitʃi]

(8) a. DELETION output-provides PALATALIZATION: $\langle x, y \rangle = \langle$ikitʃai, ikitʃi$\rangle$
   (i) $x \xrightarrow{\text{DEL.}}$ is not vacuous: /ikitʃai/ $\xrightarrow{\text{PAL.}}$ [ikitʃi]
   (ii) $\xrightarrow{\text{PAL.}} x$ is vacuous: /ikitʃai/ $\xrightarrow{\text{PAL.}}$ [ikitʃai]
   (iii) $\xrightarrow{\text{PAL.}} y$ is not vacuous: /ikiti/ $\xrightarrow{\text{PAL.}}$ [ikitʃi]

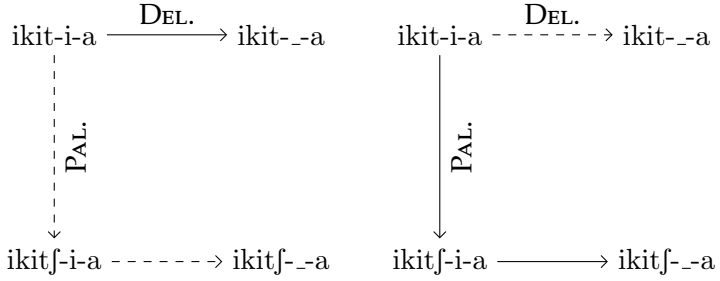
The deletion process, which precedes the palatalization process in the feeding interaction is both input-providing and output-providing: after applying deletion, the conditioning environment for palatalization is met, and the process is able to faithfully apply. As a consequence, several lexicon-grammar pairs become available that retain consistency with the data. In the case of the feeding OR [ikitʃi], one can posit the UR /ikit-_-i/ and learn the intended deletion and palatalization processes in order to capture the form. However, since deletion input-provides palatalization, the output of deletion can also be used as a compatible UR /ikit-_-i/ without having to sacrifice the learning of the palatalization process. Moreover, as deletion is also output-providing, the model can posit entirely unique

URs such as /ikitʃ-a-i/, where the palatalization process is encoded in the lexicon instead, again without losing the ability to posit the palatalization process. In other words, input and output provision not only increase the number of unique URs that can be posited, but moreover the number of grammars that are compatible with those URs. Outputs that are not as restrictive are consequently compatible with more joint hypotheses.

In contrast, the counter-feeding interaction has the deletion process follow palatalization, which means that the properties that benefited the generation of the feeding form now serves as a detriment to the counter-feeding form. Because deletion necessarily input-provides palatalization, the output of the deletion process [ikit-_-i] cannot be used as a valid UR unless palatalization is explicitly not learned. Likewise, as deletion output-provides palatalization, and as the counter-feeding form does not contain the palatalized consonant, there is no alternative, unique UR hypothesis that can be posited to generate the counter-feeding output. As a result, there are fewer compatible lexicons and grammars tied to the counter-feeding output than the feeding output. Outputs that are more restrictive are thus, unsurprisingly, compatible with fewer joint hypotheses.

Let us move onto the bleeding and counter-bleeding interactions. I enumerate the space of consistent hypotheses for the interacting ORs for each the languages in (9).

(9) Space of compatible joint lexicons and grammars for the bleeding OR [ikita] (LEFT) and counter-bleeding OR [ikitʃa] (RIGHT)



Empirically, across the palatalizing and interacting trials, both participants in the experiment and the computational model performed better when trained on the counter-bleeding language than on the bleeding language. Note, however, that the magnitude of the differences is smaller in this case than with the feeding and counter-feeding languages. The source of the asymmetry again lies in their respective surface informativity. I provide sample form pairs that demonstrate the relevant properties of each input and output interaction for the bleeding and counter-bleeding ORs in (10) and (11).

(10) a. DELETION input-removes PALATALIZATION: $\langle x, y \rangle = \langle \text{ikitia}, \text{ikita} \rangle$
   (i) $x \xrightarrow{\text{DEL.}}$ is not vacuous: /ikitia/ $\xrightarrow{\text{DEL.}}$ [ikita]
   (ii) $x \xrightarrow{\text{PAL.}}$ is not vacuous: /ikitia/ $\xrightarrow{\text{PAL.}}$ [ikitʃia]
   (iii) $y \xrightarrow{\text{PAL.}}$ is vacuous: /ikita/ $\xrightarrow{\text{PAL.}}$ [ikita]

(11) a. DELETION output-removes PALATALIZATION: $\langle x, y \rangle = \langle \text{ikitʃia}, \text{ikitʃa} \rangle$
   (i) $x \xrightarrow{\text{DEL.}}$ is not vacuous: /ikitʃia/ $\xrightarrow{\text{PAL.}}$ [ikitʃa]
   (ii) $\xrightarrow{\text{PAL.}} x$ is not vacuous: /ikitʃia/ $\xrightarrow{\text{PAL.}}$ [ikitʃia]
   (iii) $\xrightarrow{\text{PAL.}} y$ is vacuous: /ikitʃa/ $\xrightarrow{\text{PAL.}}$ [ikitʃa]

The deletion process, which again precedes palatalization in the bleeding interaction here, is input-removing and output-removing, which evokes different properties with respect to the space of compatible lexicons and grammars. Because deletion input-removes palatalization, the output of deletion does not result in a form distinct from the bleeding OR. This in turn reduces the number of unique URs compatible with the bleeding form. Thus, the number of lexicon-grammar pairs available for the bleeding form are in turn smaller than those available to the feeding form. However, because the output of deletion in this case does not create an environment for palatalization to apply, this form can be posited as a UR while still allowing the grammar to learn palatalization. This results in the bleeding OR having more compatible lexicon-grammar pairs than the counter-feeding OR.

Reversing the order of the rules results in the counter-bleeding interaction, wherein the deletion process now follows palatalization. Since the deletion process is output-removing, the counter-bleeding OR [ikitʃa] is compatible with intermediate forms generated from the output of palatalization not available to the bleeding form, e.g. /ikitʃ-i-a/. Like with the feeding and bleeding languages, each of these URs are compatible with a number of grammar hypotheses not available to the counter-feeding form. Crucially, however, because the deletion process is not output-providing given the counter-bleeding OR, additional alternative URs available to the feeding form are not available in the case of the counter-bleeding form. This results in a space of hypotheses that is larger than both the bleeding and counter-feeding form, but smaller than the feeding form. Because the bleeding form is compatible with more grammars, and the counter-bleeding form is compatible with fewer URs, however, the magnitude in the difference between the two is smaller than what is observed between the feeding and counter-feeding forms.

Interestingly the space of compatible lexicons and grammars appears to largely align with the joint predictions under the Maximum Utilization and Transparency biases. Tables comparing the predictions of the biases and the actual space of compatible URs and grammars are given in Table 3-7. This seems to suggest that the space of lexicons and grammars

Table 3-7: Comparison of the joint predictions of the Maximum Utilization and Transparency biases (TOP) versus the space of compatible lexicons and grammars (BOTTOM) for each of the four process interactions. Darker shaded cells correspond to languages predicted to be relatively more difficult to learn or consistent with fewer hypotheses, respectively.

BIASES	Feeding	Counter-bleeding	Bleeding	Counter-feeding
SPACES	Feeding	Counter-bleeding	Bleeding	Counter-feeding

indirectly encodes the Transparency and Maximum Utilization biases as a consequence of the compatible hypotheses associated with the outputs of each process interaction.

The model parameterized under moderate noise now incorporates the effect of the noisy channel on the model's performance; it illustrates what predictions the model would make if it was also allowed to consider hypotheses that nearly capture the ORs given the data. We see under this parameterization that the model produces the same qualitative observations made under the experiment: in the palatalizing trial, the model performed better when trained on the maximally-utilizing feeding and counter-bleeding interactions, while in the interacting trials, the the model performed better when trained on the transparent bleeding and feeding interactions. I will argue that this emerges as a consequence of the relative number and weight of lexicon-grammar hypotheses consistent with an alternative, but nearly identical, pattern to the one being reproduced, exhibiting properties both of similarity, as proposed by Rafferty et al. (2013), as well as uniformity (King 1969). As the main difference across the two parameterizations is the asymmetry observed between the bleeding and counter-bleeding languages, I focus our attention there.

Consider the hypothesis in (12). This hypothesis encodes an underlying /tʃ/ as well as two phonological processes: a generalized depalatalization process and a vowel deletion process. I will refer to the output of this hypothesis as the LEVELING PATTERN.

(12) Sample alternative hypothesis for the bleeding language

UNDERLYING FORM	/ikitʃ/	/ikitʃ-a/	/ikitʃ-i/	/ikitʃ-i-a/
tʃ → t	–	[ikita]	[ikiti]	[ikatia]
V → ∅ / _V	[ikit]	–	–	[ikata]
EXPECTED FORM	[ikit]	[ikita]	[ikiti]	[ikata]

The existence of the generalized depalatalization process ensures that no [tʃ] will ever surface in the output. The output distribution that emerges from this hypothesis is identical in all cases to the original bleeding language except in the palatalizing trial, where each

predicts [ikiti] versus [ikitʃi], respectively. Under a parameterization with moderate noise, the cost of getting one form wrong is not as costly. If there are sufficient hypotheses associated with this alternative surface distribution over the pattern actually observed, the collective weighted posterior over these nearly consistent hypotheses may be able to overpower the contribution the consistent hypotheses. It is not difficult to intuit that the number of lexicons and grammars that can generate the leveling pattern is vast; so long as the palatalization processes are underlearned, or if they are followed at some point with the generalized depalatalization process, the leveling pattern will emerge. If this were to produce a meaningful effect, we expect the model to have relatively lower accuracy in the palatalizing trials under a high-noise parameterization compared to a low-noise parameterization. This is exactly what we see.

In contrast, consider the hypothesis in (13). This hypothesis encodes an underlying /tʃ/ as well as two phonological processes: a contextual depalatalization process and a vowel deletion process. I will refer to the output of this hypothesis the DEPALATALIZATION PATTERN.

(13) Sample alternative hypothesis for the counter-bleeding language

UNDERLYING FORM	/ikitʃ/	/ikitʃ-a/	/ikitʃ-i/	/ikitʃ-i-a/
tʃ → t / _{C, #}	–	–	–	–
V → ∅ / _V	[ikit]	–	–	–
EXPECTED FORM	[ikit]	[ikitʃa]	[ikitʃi]	[ikatʃa]

The output distribution that results from this lexicon-grammar hypothesis is one in which the palatalized consonant emerges in all contexts except in the faithful trials, predicting [ikitʃa] instead of the expected [ikitʃi]. As it was in the case above, under a parameterization with moderate noise, the cost of getting the one form wrong is relatively uncostly. However, unlike in the bleeding language, the expected shift in performance given this alternative pattern is not met. Instead, we primarily observe a decrease in performance on the interacting trials under moderate noise. Unlike the leveling pattern in (12), the depalatalization pattern in (13) is not associated with as many grammars, requiring strictly a generalized palatalization process followed by contextual depalatalization, or contextual depalatalization given underlying /tʃ/. As a result, the relative contribution of these hypotheses on the overall performance of each form is lower. Instead, the drop in performance emerges from the combined weighted probability assigned to the leveling pattern. While the relative posteriors assigned to each individual hypothesis is lower than in the bleeding language, the large number of hypotheses in the space associated with this pattern still contributes

to the predictions of the model.

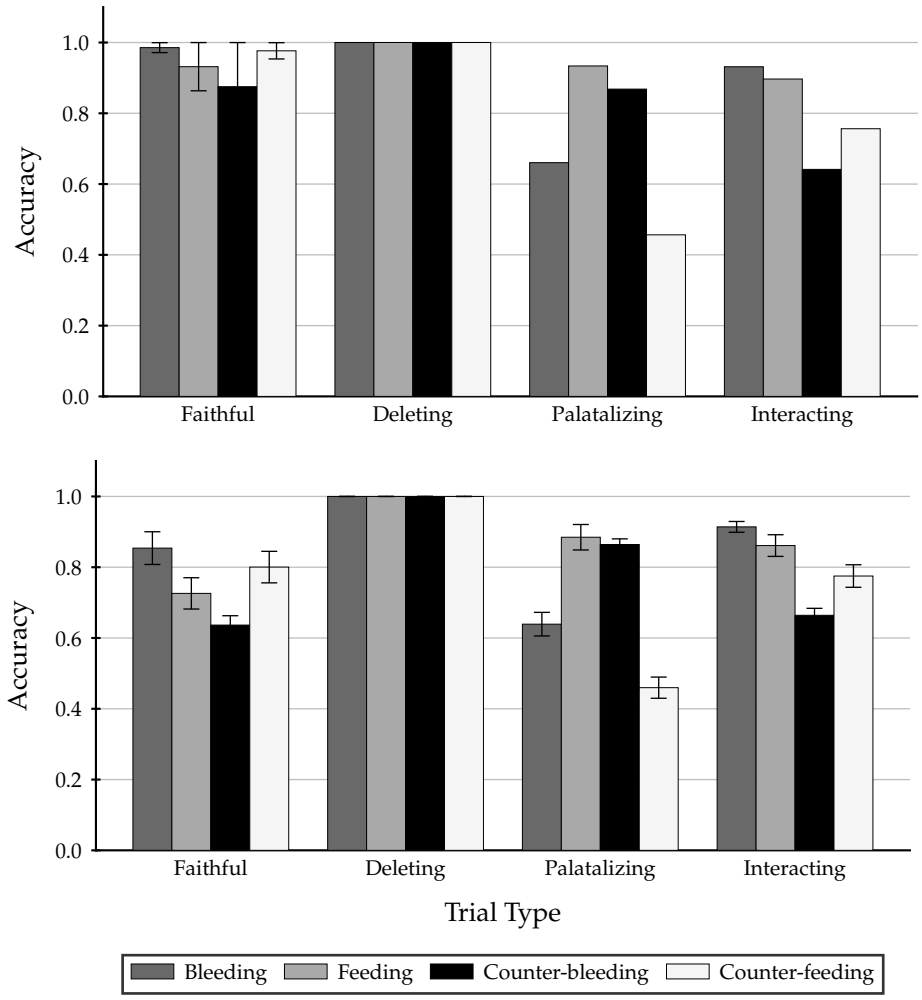
The discussion here is compatible with the mislearning hypothesis presented earlier, although the manner in which the noisy-channel achieves these results is slightly different. While previous approaches get this result as a consequence of incomplete learning, the noisy-channel model instead formalizes this relationship in terms of grammar compatibility and similarity. Some distributions are compatible with more lexicons and grammars than others, allowing it to more easily land on a hypothesis that reproduces the data. For example, the feeding language outperformed the counter-feeding language because the former has more compatible analyses that retains the original phonological generalizations compared to the latter. In addition, certain distributions are more similar to alternative distributions, resulting in ambiguity of what the correct pattern to optimize the model for is. This ambiguity results in the model producing certain forms well at the detriment of others. For example, the model when trained on the bleeding language performed better than when trained on the counter-bleeding language in the interacting trials but worse in the palatalizing trials as it was confused between the observed distribution and the alternative, leveling distribution. The noisy-channel model achieves these results with minimal assumptions over the lexicon and no assumptions over the search process.

The results here indicate that patterns with similar alternative distributions are more likely to be mislearned than those with less similar ones. Moreover, regardless of similarity, patterns that are *a priori* more plentiful in the hypothesis space are still capable of influencing the predictions of the model, if the number is sufficiently large. This property roughly corresponds to the insights made by Rafferty et al. (2013), who found that the persistence of a pattern corresponds to the similarity of that pattern to alternatives, as well as the relative space that pattern occupies over the set of all possible surface patterns. Interestingly, patterns with the largest *a priori* space typically correlate with those in which the distribution over ORs is more uniform. This property roughly correponds to the general principle behind the Uniformity bias, as patterns with relatively less evidence of a particular alternant is compelled to dropp the alternation entirely.

Multiple paradigm distribution results

Does the absolute type frequency affect the results observed above? To test this, I provide the model with the data consisting of three complete and incomplete examples of each paradigm type. The results of the model are given in Figure 3-4. Like as was observed in the single paradigm distribution, the model succeeds in capturing the main empirical asymmetries made by Prickett. However, the model requires a more noisy parameterization in order to achieve the results. The reason this occurs is that the more data that is given

Figure 3-3: Average accuracy of the model for the Prickett experiment for each trial type by language. Top: single paradigm distribution, $\lambda = 3$. Bottom: multiple paradigm distribution, $\lambda = 1$.

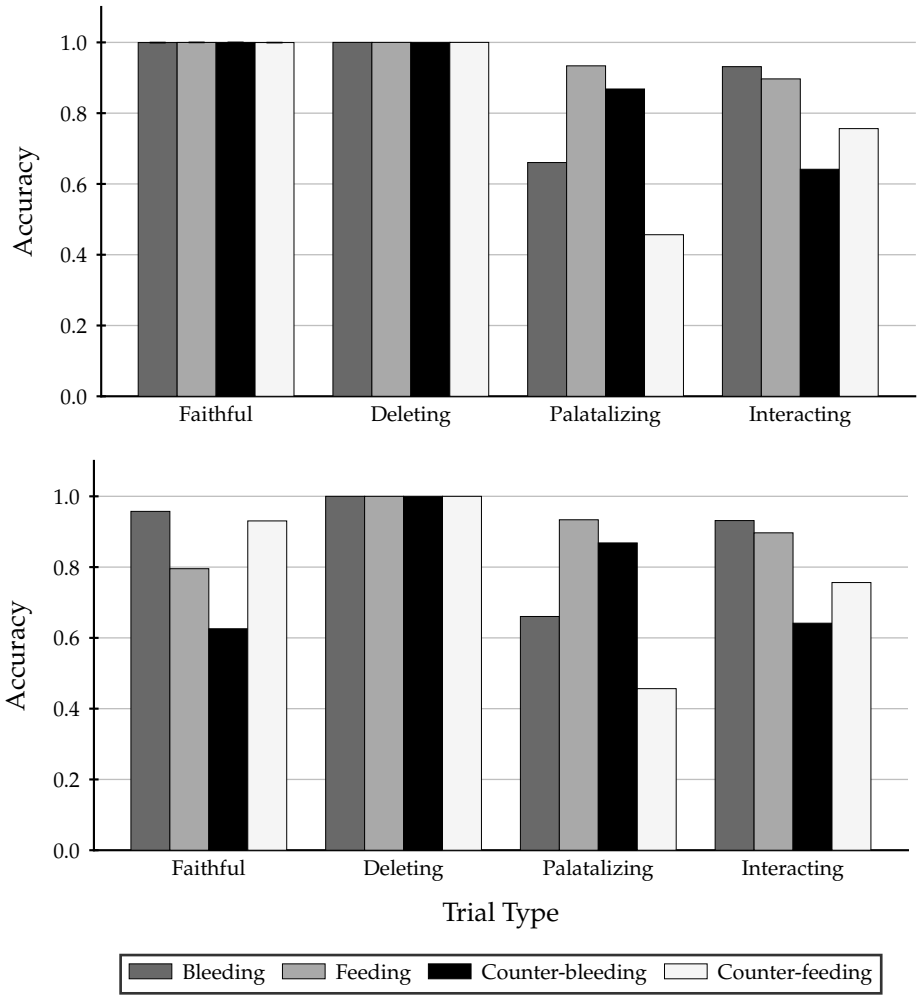


to the model, the more evidence the model has in terms of calculating the likelihood. As such, generalized deviations from the surface forms results in a greater hit to the likelihood. For example, the hypothesis in (13) would incur a lower likelihood as the absolute number of forms that are incorrect in the type distribution is greater than in the single paradigm distribution. As such, a lower λ is necessary to allow sufficient probability mass to be allocated to the nearly consistent forms.

3.2.4 Discussion

I wish to bring to attention one last observation not explicitly noted by Prickett, but predicted under the model; namely, the performance of the faithful trials by the model differed

Figure 3-4: Average accuracy of the model for the Prickett experiment for each trial type by language under the SINGLE PARADIGM distribution. TOP: the faithful trials consist only of the k-final paradigm. BOTTOM: the faithful trials consist only of the t-final paradigms.



depending on the language. The distribution follows an anti-Maximum Utilization bias, with performance improving when trained on the bleeding and counter-feeding languages over the feeding and counter-bleeding languages. Note, however, that the mean accuracy within a language has a great deal of variance.

Recall that the the faithful trial consists of multiple different paradigm slots. These include the lexical contexts in which the coronal-final stem combines with the /-a/ suffix, as well as the lexical contexts in which the velar-final stem combines with either the /-i/ or /-a/ suffix. In other words, the faithful trial aggregates over two different stems, one in which the stem alternates and one in which the stem does not. Separating the faithful trials to comprise only of the coronal-final stems or the velar-final stems nets the two plots in

Figure 3-4. As can be observed, there is no difference in performance among the velar-final stem paradigms, with the model performing at ceiling, whereas for the coronal-final stem paradigms, the shape mimics the anti-Maximum Utilization bias aforementioned. This shape is visually similar in shape to the faithful trials observed in the experimental results as well, with the mean performance being higher among participants trained on the non-maximally utilizing bleeding and counter-feeding languages. In the experiment, however, Prickett noted that the observed difference was not significant; this may have been caused by the aggregation over both paradigms.

In this section, we observed that the model was able to successfully reproduce the observed empirical asymmetry observed in the Prickett artificial language learning experiment. We found that the model was able to do so due to the difference in the collective space of lexicon-grammar pairs consistent or nearly consistent with each language. I noted, however, that a crucial component in the model's performance in capturing an alternation was the relative frequency of each allophone; the model with greater noise was able to perform better in the interacting trials and worse on the palatalizing trials when trained on the counter-feeding language because there were more instances of the non-palatalized [t] over its palatalized counterparts [tʃ]. In the next experiment, I use this knowledge in order to explain a seemingly contradictory result featuring the exact same feeding and counter-feeding interactions discussed here; the feeding OR is not always easier for the model to reproduce than the counter-feeding OR.

3.3 Modeling Kim (2012)

Kim performed a poverty-of-the-stimulus artificial grammar learning experiment to assess which phonological interaction participants preferred when given ambiguous data. To do this, Kim trained 12 participants in a single language. The language is composed of several morpho-phonological paradigms. Each paradigm consists of three lexical contexts: ① a stem of the shape CVC([a]) in isolation, ② a stem of the shape ([i])CVC in isolation, and ③ both stems followed by an /-a/ suffix. The first stem in the paradigm consists of three broad categories: the coronal-final stems, e.g. /kat/ and /kad/, the non-coronal consonant-final stems, e.g. /kak/ and /kag/, and the vowel-final stems, e.g. /kata/ and /kaka/. The second stem in the paradigm consists of two categories: ① the vowel-initial stems, e.g. /ipak/ and /ikik/, and ② the consonant-initial stems, e.g. /pak/ and /kik/. To simplify the discussion, I limit the discussion over coronal-final stems to [t] and velar-final stems to [k]. Each OR in a paradigm is generated by sequentially applying in some order the deletion and palatalization processes outlined in (14) to each UR. Each OR of the language

Table 3-8: Schematic URs and ORs for the ambiguous toy language in Kim (2012).

Feeding Language

Underlying Form	/kak/	/pap/	/kak-pap-a/	/kaki/	/apap/	/kaki-apap-a/
V → ∅ / _V	–	–	–	–	–	[kakapapa]
t → tʃ / _i	–	–	–	–	–	–
Observed Form	[kak]	[pap]	[kakpapa]	[kaki]	[apap]	[kakapapa]

Underlying Form	/kat/	/ipap/	/kat-ipap-a/	/kata/	/ipak/	/kata-ipak-a/
V → ∅ / _V	–	–	–	–	–	[katipaka]
t → tʃ / _i	–	–	[katʃipapa]	–	–	[katʃipaka]
Observed Form	[kat]	[ipap]	[katʃipapa]	[kata]	[ipak]	[katʃipaka]

Counter-feeding Language

Underlying Form	/kak/	/pap/	/kak-pap-a/	/kaki/	/apap/	/kaki-apap-a/
t → tʃ / _i	–	–	–	–	–	–
V → ∅ / _V	–	–	–	–	–	[kakapapa]
Observed Form	[kak]	[pap]	[kakpapa]	[kaki]	[apap]	[kakapapa]

Underlying Form	/kat/	/ipap/	/kat-ipap-a/	/kata/	/ipak/	/kata-ipak-a/
t → tʃ / _i	–	–	[katʃipapa]	–	–	–
V → ∅ / _V	–	–	–	–	–	[katipaka]
Observed Form	[kat]	[ipap]	[katʃipapa]	[kata]	[ipak]	[katipaka]

is generated through the sequential application of the same deletion and palatalization processes as observed in (1), which I repeat in (14).

(14) a. Deletion: V → ∅ / _V
b. Palatalization: t → tʃ / _i

Depending on which stem types combine together, the learner receives evidence of the deletion process, palatalization process, and their interaction. Though not done in original experiment, to maintain parallelism with the discussion done with Prickett, I organize the lexical contexts of each language into distinct trial types according to which process was expected to apply. When a non-coronal consonant-vowel-final stem combines with a vowel-initial stem, e.g. /kaka-ipak-a/, we produce the deleting form, e.g. /kakipaka/.

Table 3-9: Trial types and schematized choices for the toy language in Kim (2012). The rows in each sub-table correspond to the trial type, expected UR, intended response, and alternate response.

/kak/	/pap/	FAITHFUL /kak-pap-a/
		[kak-pap-a]
		*[katʃ-pap-a]

/kaki/	/apap/	DELETING /kaki-apap-a/
		[kakapapa]
		*[kakaipapa]

/kat/	/ipap/	PALATALIZING /kat-ipap-a/
		[katʃipapa]
		*[katipapa]

/kata/	/ipak/	INTERACTING /kata-ipak-a/
		?[katʃipaka]
		?[katipaka]

When a consonant-final stem combines with a consonant-initial stem, e.g. /kat-pak-a/ or /kak-pak-a/, we produce the faithful form, e.g. [katpaka] and [kakpaka]. When a coronal-final stem combines with an vowel-initial stem and /-a/ suffix, e.g. /kat-ipak-a/, we produce the palatalizing form, e.g. [katʃipaka]. Lastly, when a coronal consonant-vowel-final stem combines with a vowel-initial stem and /-a/ suffix, e.g. /kata-ipak-a/, we produce the interacting form, which differs based on the order of rules learned. If the feeding interaction is learned, the grammar produces an ER containing the palatalized consonant, e.g. [katʃipaka], whereas if the counter-feeding interaction is learned, the grammar produces the ER containing the non-palatalized consonant, e.g. [katipaka]. A representation of each relevant trial type and respective form is shown in Table 3-8. We see that each trial type is isolated to its own paradigm. Moving forward, I will refer to each paradigm according to the trial type it is associated with, e.g. faithful paradigm, deleting paradigm, and so on.[5]

Like Prickett, Kim organized the experiment into two phases: a training phase and a testing phase. The training phase consisted of two different kinds of trials: listening trials and feedback trials. In the listening trials, participants were given each stem in isolation to the speaker, paired with an image corresponding to their respective meanings. They were then given the complex form and the intended response for that form via both an audio recording of the pronunciation as well as the orthography. For example, in the faithful trial, the participant hears the ORs [kat] and [pap], paired with an image of the relevant, respective meanings. They are then given an image referring to the complex form, paired

[5] The original experiment does not provide any description of trial types nor explicit description of the actual segments used to generate the stems and paradigms of the language. These are reconstructions based on good-faith interpretations. All errors are my own.

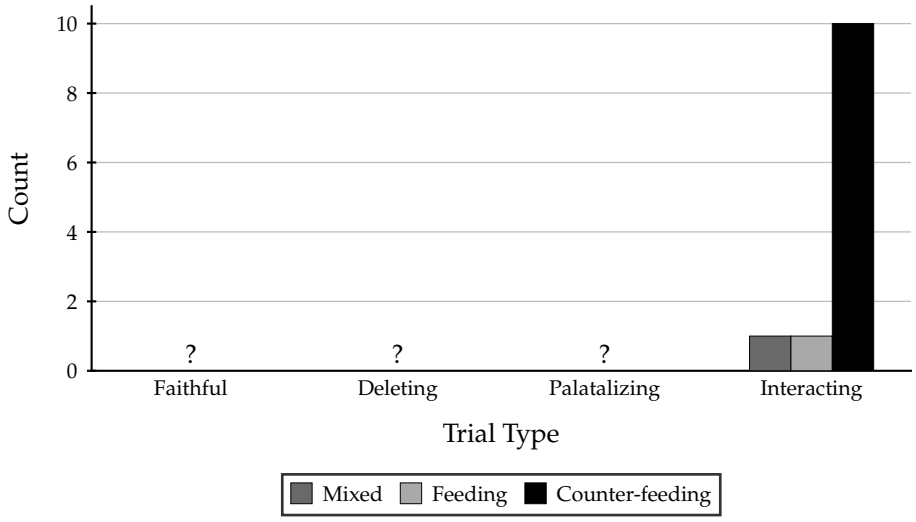
with an audio recording of that form [katpapa] as well its orthographic representation *"katpapa."* In the feedback trials, participants are given each stem in isolation, then asked to produce the complex form by combining each stem with the /-a/ suffix. For example, in the palatalizing trial, the participant first hears the forms [kat] and [ipap], paired with its correspondent image. The participant is then provided an image for the complex lexical context and ask to generate what they believe is the form consistent with the language. Feedback in the form of an audio cue of the expected pronunciation, e.g. [katʃipapa], and an orthographic representation of the word, e.g. *"kachipapa"*, is then given to the participant. Participants were given examples of the faithful trials, deleting trials, and palatalizing trials, but were given no examples of the interacting trials. In the testing phase, participants were given exclusively feedback trials. Participants were presented novel forms testing the palatalization and deletion processes in isolation, much like in the training phase. However, in addition to these forms, participants were also given paradigms in which the two learned phonological processes potentially interact; for example, the lexical context combining /pata/ and /ipak/ with the suffix /-a/. The goal was to determine whether participants had a preference for the feeding or counter-feeding form in the absence of explicit information of either's existence. Schematizations of the intended and alternate responses under a feeding versus counter-feeding grammar are given in Table 3-9.[6] The intended response for the interacting trial corresponds to the feeding OR, e.g. [katʃipaka], as it is the one expected under both the Transparency and Maximum Utilization biases. The alternate response for the interacting trial thus corresponds to the counter-feeding OR, e.g. [katipaka]. In the training phase, participants were exposed to 44 faithful paradigms, 8 deletion paradigms, and 8 palatalization paradigms. In the testing phase, participants were given 4 faithful paradigms, 6 deleting paradigms, 6 palatalizing paradigms, and 12 interacting paradigms.

3.3.1 Experimental results

Kim grouped participants based on which form individuals chose in the interacting trials; for example, given /kata-ipak-a/, did the participant tend to select the feeding ER, i.e. [katʃipaka], counter-feeding ER, e.g. [patipaka], or did they freely vary between the two forms? The results of the experiment are shown in Figure 3-5. Kim found that participants

[6] It is not clear whether the experiment was performed as a free-response or as a forced-choice experiment. I opt to interpret the experiment as a forced-choice experiment for two reasons. First, it provides us a closer level of parallelism with the Prickett results, which allow us to better compare the results observed in each. Second, renormalizing the responses to only include the intended and alternate responses will make the differences between the feeding and counter-feeding response rates more clear. Since we do not know the true identity of the alternate responses, I adopt the same alternatives used by Prickett.

Figure 3-5: Counts of participants based on response rates in the interacting trials in Kim (2012).



overwhelmingly preferred to produce the counter-feeding form over the feeding form; of the twelve participants trained and tested, ten preferred the counter-feeding form, only one preferred the feeding form, and one exhibited free-variation between the two.

These results are particularly interesting for three reasons. First, although the individual processes involved in this experiment are identical to those observed in Prickett's experiment above, the distribution of forms given to the participant vary greatly between the two; while Prickett kept the relative frequency uniform over all trial types, Kim provided participants with many more examples of faithful trials over the deleting and palatalizing trials, as well as no examples of the interacting form at all. Second, the space of candidate grammars differs between the two distributions. As the V~ ∅ occurs word-internally in the Kim experiment, alternative analyses such as interconsonantal i-insertion can now be reasonably entertained to capture the data. This is in contrast to the Prickett experiment, in which the alternation occurs word-finally; thus, the environment for i-insertion is not available and therefore not entertained. Lastly, and perhaps most striking, is that the results observed here appear to be in direct opposition to what was found in Prickett (2019) and what was modelled under the lexicophonological learner. Where Prickett found that participants trained on the feeding interaction performed statistically better across-the-board over those trained on the counter-feeding interaction, Kim found the opposite preference, with participants seemingly having a much higher preference for the counter-feeding form over the feeding form.

I will argue that the source of the seemingly contradictory results is the difference in the distribution of forms given in each experiment. As I discussed in the previous section, the

Table 3-10: Inventory used to model the Kim (2012) data.

	[cons]	[lab]	[cor]	[dist]	[high]
p	1	1	-1	0	0
t	1	-1	1	-1	0
tʃ	1	-1	1	1	0
k	1	-1	-1	0	1
a	-1	1	-1	0	-1
i	-1	1	-1	0	1

results of the learner are dependent on the space of both the consistent and nearly consistent lexicon and grammar pairs. I will argue that the effect of this distribution on the space of nearly consistent hypotheses serves as the primary catalyst for the reverse asymmetry: the ORs are organized in such a way that the model is exposed to the non-palatalized [t] much more frequently than its palatalized counterpart. The noisy channel thus pushes the model to allocate more probability mass to the incorrect, but nearly correct hypotheses in which the alternation is eliminated from the surface.

3.3.2 Computational overview

In this section, I seek to investigate whether providing the model with data skewed towards faithful data affects the model's ability to learn the underlying pattern. In order to evaluate the performance of the model on this experiment, we must provide the model with the hypothesis space over the lexicon and the rules, as well as the set of data.

Hypothesis space over lexicon

The inventory given to the model is provided in Table 3-10. This inventory consists of 6 segments and 5 features[7]. The prototype and contextual URs consist of any possible permutation of these segments.[8]

Hypothesis space over rules

Next, I lay out the space of possible rules in (15). The rule space consists of two vowel alternation processes: one in which a vowel deletes before another vowel, and one in

[7] The difference in the number of segments between this experiment and the Prickett experiment has no effect on the main empirical observations I will make with the model. While the difference in features may have an impact on the computation of the edit distances, as our analysis does not rely on the feature inventory, I do not expect this to make a significant difference on the results.

[8] Again, I set the upper bound of UR lengths to be 5: $\textsc{max}(|u_{x0}|) = \textsc{max}(|u_{xc}|) = 5$.

Table 3-11: Schematic dataset for the Kim language given to the model.

kak	$\langle \mu_1 \rangle$	pap	$\langle \mu_2 \rangle$	kakpapa	$\langle \mu_1, \mu_2, \mu_A \rangle$
kaki	$\langle \mu_3 \rangle$	apap	$\langle \mu_4 \rangle$	kakapapa	$\langle \mu_3, \mu_4, \mu_A \rangle$
kat	$\langle \mu_5 \rangle$	ipap	$\langle \mu_6 \rangle$	katʃikaka	$\langle \mu_5, \mu_6, \mu_A \rangle$
kap	$\langle \mu_7 \rangle$	pak	$\langle \mu_8 \rangle$	???	$\langle \mu_7, \mu_8, \mu_A \rangle$
kapi	$\langle \mu_9 \rangle$	akak	$\langle \mu_{10} \rangle$	???	$\langle \mu_9, \mu_{10}, \mu_A \rangle$
pat	$\langle \mu_{11} \rangle$	ikak	$\langle \mu_{12} \rangle$	???	$\langle \mu_{11}, \mu_{12}, \mu_A \rangle$
pata	$\langle \mu_{13} \rangle$	ipak	$\langle \mu_{14} \rangle$	???	$\langle \mu_{13}, \mu_{14}, \mu_A \rangle$

which a vowel is inserted between two consonants. Morever, the rule space includes both generalized and contextual palatalization and depalatalization processes.

(15) a. Vowel Deletion: V → ∅ / _V
b. Vowel Insertion: ∅ → i / C_C
c. Palatalization: t → tʃ / _i
d. Generalized Palatalization: t → tʃ
e. Depalatalization: tʃ → t / _{C, #}
f. Generalized Depalatalization: tʃ → t

Distribution of the data

The model was provided with the schematized dataset in Table 3-11. The data consists of both complete and incomplete paradigms for the faithful, deleting, and palatalizing paradigms. This is to ensure the model has independent evidence of each individual trial, as well as the identity of each noun in isolation, mimicking the training phase of the experiment. The incomplete paradigms consist only of each stem in isolation, with the context in which both are combined with the /-a/ suffix being held-out. In addition, I provide the model only incomplete paradigms for the interacting paradigms. This is to evaluate whether model has a preference for the counter-feeding or feeding form, mimicking the testing phase and goal of the experiment.

In order to test whether the distibution of the data truly has an effect on the results of the model, I evaluate the learner under two distributions: the uniform distribution and the skewed distribution. Under the uniform distribution, the model is presented only one complete paradigm for each paradigm type, excluding the interacting paradigm. This corresponds exactly to the dataset outlined in Table 3-11. Under the skewed distribution, the type frequencies of each paradigm are adjusted to roughly match those observed in

Table 3-12: Type distributions for modelling the Kim (2012) experiment.

	Uniform		Skewed	
	Observed	Held-out	Observed	Held-out
Faithful	1	1	10	1
Deleting	1	1	2	2
Palatalizing	1	1	2	2
Interacting	1	1	0	4

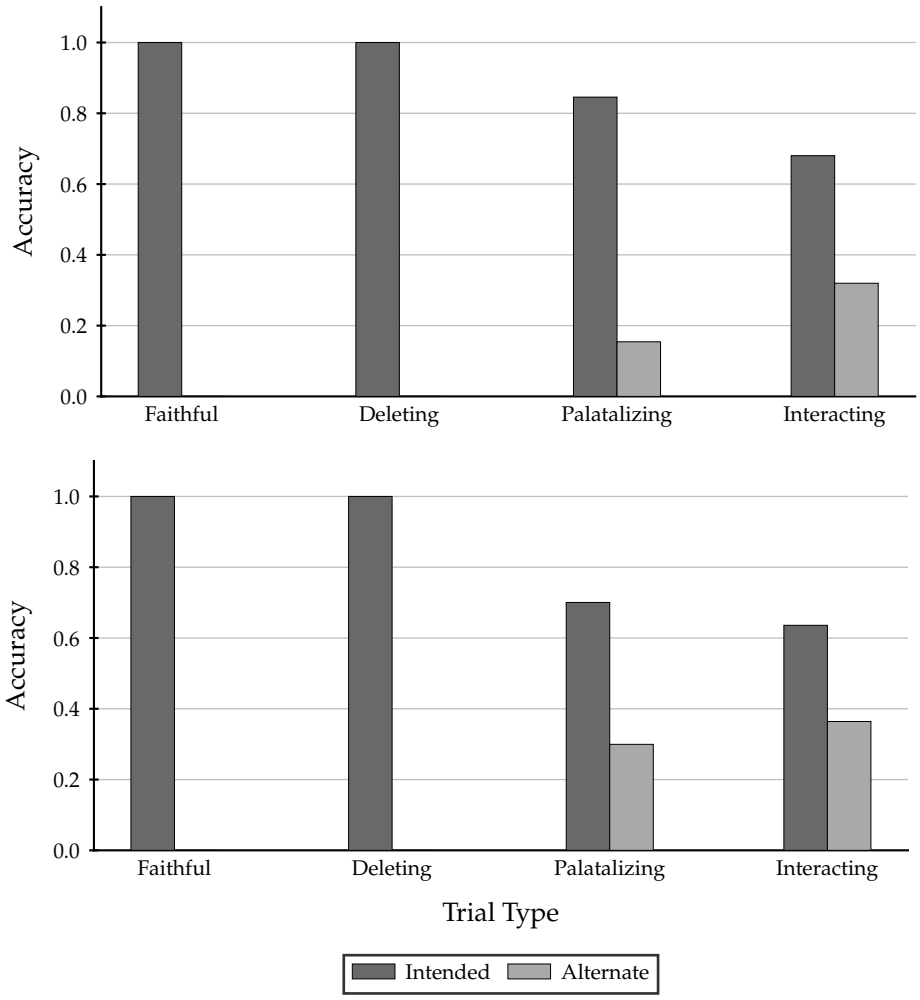
the experiment. This is typified in Table 3-12. Again, due to technical reasons, we must truncate the total number of forms given to the model. However, I cannot cleanly maintain the relative type frequencies while keeping the total number of lexical contexts given to the model low, due to the ratio of forms being extremely skewed. As the main point of the investigation is simply to examine whether the additional of more [t] forms affects the model's ability to detect the alternation, small differences in the relative frequencies will not prevent us from evaluating this. Lastly, I do not know the actual distribution of faithful paradigms. In the absence of overt data, I will assume that the faithful paradigms are equally distributed between coronal-final stems, e.g. [kat], and non-coronal consonant-final stems, e.g. [kak].

Like before, I will be assessing the model's performance based on what predictions the model makes on held-out data. I compute the accuracy as observed in the previous experiment, where I renormalize the predictive probability to only include the probabilities of generating the intended and alternate option in the forced-choice task:

$$\text{(16)} \qquad \textsc{Accuracy} = \frac{P(e^*_{\textsc{intended}} \mid o)}{P(e^*_{\textsc{intended}} \mid o) + P(e^*_{\textsc{alternate}} \mid o)}$$

How do we relate the performance of the model to the counts observed in the experiment? We are not given information regarding the actual performance within a trial type, nor information on how the participants were categorized into the mixed, feeding, and counter-feeding groups; for example, what is the minimum frequency a participant must select the counter-feeding option [katipaka] to be categorized into the counter-feeding groups? Moreover, as discussed above, the experiment did not report the performance of the model on the faithful, deleting, or palatalizing trials; as such, we do not know whether participants who favored the counter-feeding form truly learned the palatalization process or not. In order to compare the results of the model to the results of the experiment, then, I adopt to take a qualitative approach. The metric I will use to evaluate the model's performance is to determine whether the model can minimally replicate the preference

Figure 3-6: Average accuracy of the model for the Kim experiment for each trial type under the UNIFORM distribution. TOP: $\lambda = 10$. BOTTOM: $\lambda = 3$.



for the counter-feeding option: if the model assigns a higher relative posterior predictive probability to the counter-feeding form over the feeding form, I will interpret that the model makes the same predictions as observed in the experiment.

3.3.3 Computational results

I examine the results of the model for each type distribution in this section. This section is oganized into two parts, corresponding to the uniform and skewed distributions discussed in Table 3-12. The results of the model for each distribution are discussed in the following subsections.

Uniform distribution results

I tested the model under two parameter settings over the noisy channel: one with ① low noise $\lambda = 10$ and ② moderate noise $\lambda = 3$. We see that under both parameterizations, the model fails to capture the observed empirical asymmetries; the model strongly prefers the feeding form over the counter-feeding form.

The source of these results mirrors the reasoning given in the previous section: feeding interactions generate distributions that constitute a larger portion of the weighted inferred lexicophonological space than counter-feeding interactions. Increasing the noise of the model does have an effect on the performance of the model by increasing the overall probability space nearly consistent hypotheses, such as the leveling pattern given above, occupy. There are more ORs containing the segment [t] than the segment [tʃ] in the data; as such, nearly consistent hypotheses in which the alternation is dropped in favor of only generating [t] are given relatively more of the posterior space, pushing more of the predictive probability towards the counter-feeding form. This unfortunately is not nearly sufficient to overcome the overall preference for the feeding form.

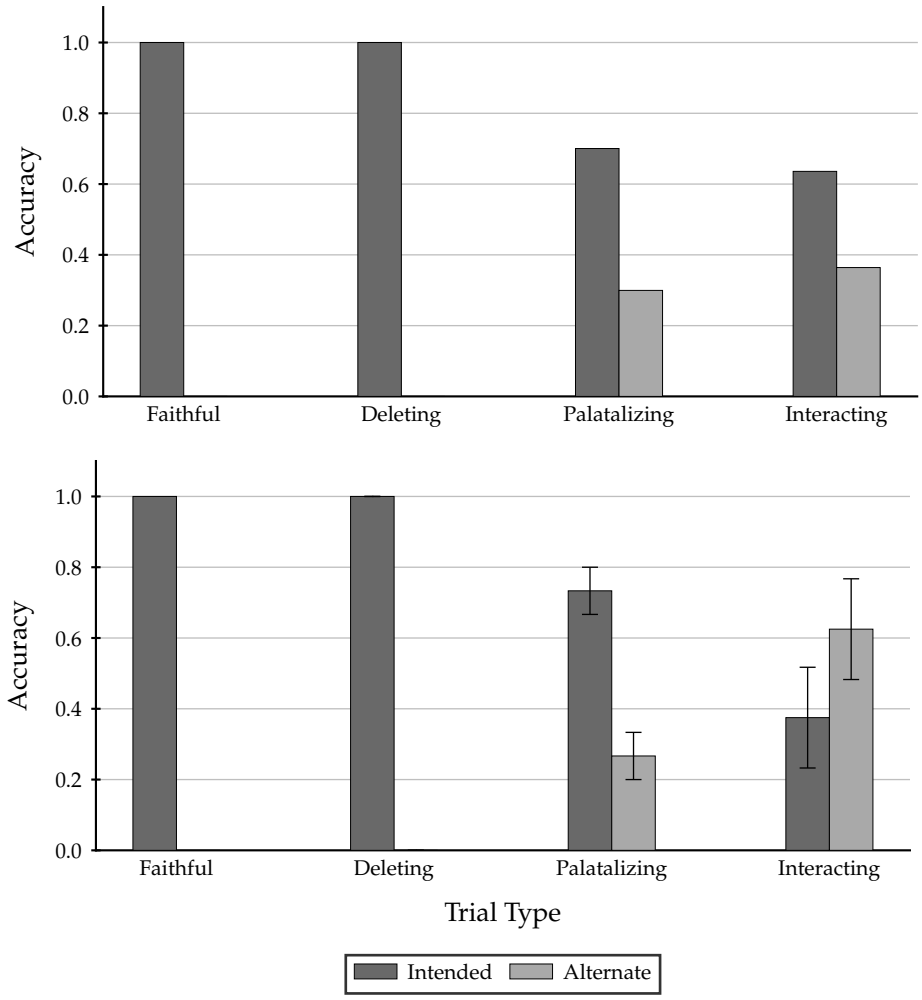
However, the uniform distribution does not reflect the true distribution of ORs given in the experiment; as presented earlier, the participants of the experiment were given many more examples of the faithful paradigm than either of the individual process. I argue that it is this difference that acts as the crucial component to achieving the results under the model.

Skewed distribution results

Is giving the model more examples of the non-palatalized consonant [t] sufficient to attain the observed experimental results? To test this, I provide the model with the skewed distribution given in Table 3-12. The results of the model are given in Figure 3-7. We see that given the set of data and parameterization, the model indeed succeeds in capturing the main empirical asymmetry observed by Kim.

The story is identical to what was observed in Prickett: while the number of consistent hypotheses is greater for the feeding form over the counter-feeding form, the number of nearly consistent grammars overpower this space. This is because the prevalence of non-alternating [t] forms and a greater amount of noise allows the model to assign more probabilities to hypotheses that nearly capture the data by underlearning surface palatalization. These results appear to indicate that the learnability of process interactions is affected not only by the difference in distributions generated from the process interaction, but moreover how many of each alternant is given to the model, and thus the learner.

Figure 3-7: Average accuracy of the model for the Kim experiment for each trial type. Top: uniform distribution, $\lambda = 3$. Bottom: skewed distribution, $\lambda = 1$.



3.3.4 Discussion

Before closing off this section, I wish to explore in a bit more detail the consequences of accepting this interpretation of the Kim results. While the distribution of the data is enough to push the model to prefer the counter-feeding interacting form over the feeding interacting form, the model does as precisely by *not* learning an interaction at all; rather, it loses the palatalization process entirely. Interestingly, the performance on the palatalizing trial does drop when increasing the level of noise in the model, but still has a much higher predictive probability than the non-palatalized form ~ 75% versus ~ 25% for the palatalized and non-palatalized option, respectively. Kim notes that the palatalization process and deletion process were tested independently in the experiment, but the results were not reported.

We expect that if palatalization amongst the participants was particularly low, those results would have been mentioned. As such, it can also mean that the model *fails* to capture the empirical observations of the experiment. With the absence of information, unfortunately, we cannot tell. Regardless, it illustrates the effect the relative distribution of forms has on the performance the model has on held-out forms; having greater examples of one alternant in the data in an alternation will direct the model away from learning the alternation entirely.

3.4 Conclusion

In this chapter, I explored the predictions of the model on two different AGL experiments, both of which were generated using the same underlying process interaction: a deletion process and a palatalization process. However, not only do the distributions and conclusions of the two experiments vary wildly, they also assume the learning of the underlying forms in modelling the results. I illustrate that, given few assumptions, a model that jointly learns the URs and mappings is also able to capture the same empirical asymmetries by way of the space of consistent and nearly consistent lexicons and grammars that can generate the language. In the next chapter, I investigate in more detail how the individual contributions of the space of consistent versus nearly consistent hypotheses affect the results of the model.

Chapter 4

The Data, Lexicon, and Effect on the Lexicophonological Space

In this dissertation, I proposed that the generalizability of a pattern – particularly, patterns generated from different process interactions – is dictated primarily by the difference in the inferred lexicophonological spaces. This space is influenced in part by the distribution of the data, as well as the space of lexicons. The aim of this chapter is to examine more rigorously the effect of each on the model's performance. In the following sections, I perform several exploratory computational experiments aimed at probing the effect of the data and lexicon on the ability of the noisy-channel model to learn and extend the generalizations of a pattern to held-out forms. I land on two important findings: ① the inferred space of lexicons and grammars is primarily shaped by the relative frequencies of each alternant in an alternation, but can occasionally be overcome when the informativity of the distribution is particularly restrictive, and ② despite allowing the model to encode surface alternations as part of the lexicon rather than the grammar, under a noisy parameterization, the model rarely applies this strategy, instead capturing most alternations through the productive application of the grammar. Ultimately, both converge on the same overall conclusion: the space of nearly consistent hypotheses serves as the main driving force behind predicting the direction of asymmetry, consistent with the findings made by Rafferty and collegues (2013). For the rest of the chapter, I will use the revised Baković languages as the baseline.

4.1 Experiment 1: How the Data Shapes Generalization

I have demonstrated in the previous chapter that the data has a considerable impact on the inferred weighted lexicophonological space, which in turn influences the reproduceability

Table 4-1: Sample feeding and counter-feeding languages from Baković (2011).

Feeding Language

ikit	$\langle \mu_1 \rangle$	ikita	$\langle \mu_1, \mu_A \rangle$	ikitʃi	$\langle \mu_1, \mu_B \rangle$	ikiti	$\langle \mu_1, \mu_A, \mu_B \rangle$
ikik	$\langle \mu_2 \rangle$	ikika	$\langle \mu_2, \mu_A \rangle$	ikiki	$\langle \mu_2, \mu_B \rangle$	ikiki	$\langle \mu_2, \mu_A, \mu_B \rangle$
akit	$\langle \mu_3 \rangle$	???	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_3, \mu_B \rangle$	???	$\langle \mu_3, \mu_A, \mu_B \rangle$
akik	$\langle \mu_4 \rangle$	???	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_4, \mu_B \rangle$	???	$\langle \mu_4, \mu_A, \mu_B \rangle$

Counter-feeding Language

ikit	$\langle \mu_1 \rangle$	ikita	$\langle \mu_1, \mu_A \rangle$	ikitʃi	$\langle \mu_1, \mu_B \rangle$	ikiti	$\langle \mu_1, \mu_A, \mu_B \rangle$
ikik	$\langle \mu_2 \rangle$	ikika	$\langle \mu_2, \mu_A \rangle$	ikiki	$\langle \mu_2, \mu_B \rangle$	ikiki	$\langle \mu_2, \mu_A, \mu_B \rangle$
akit	$\langle \mu_3 \rangle$	???	$\langle \mu_3, \mu_A \rangle$	???	$\langle \mu_3, \mu_B \rangle$	???	$\langle \mu_3, \mu_A, \mu_B \rangle$
akik	$\langle \mu_4 \rangle$	???	$\langle \mu_4, \mu_A \rangle$	???	$\langle \mu_4, \mu_B \rangle$	???	$\langle \mu_4, \mu_A, \mu_B \rangle$

of the forms within a pattern. There are two key differences in the inferred hypothesis spaces associated with each language: ① the number of lexicons and grammars compatible with the data, and ② the number of lexicons and grammars that were nearly compatible with the data. I argued that the distribution of consistent and nearly consistent hypotheses is influenced by two emergent, interacting properties: ① the informativity of the distribution, and ② the relative number of each alternant found in the distribution. In this section, I verify this claim and investigate the interaction of these two aspects of the data by performing an exploratory computational experiment. The section is organized in the following manner. First, I provide the overview and motivation behind the experiment by explicating the distinction between the two properties and their respective contribution to the lexicophonological space. Second, I introduce the manipulation over the data that will be used in order to distinguish the effects of each property. Finally, I present the computational outcomes and provide an exploratory analysis of the results.

4.1.1 Overview and motivation

The effect that emerges from the distribution of ORs given to the learner is different from the effect that emerges from the difference in relative frequencies of each alternant within an alternation. I use the revised Baković languages presented in Chapter 1 to elaborate. I repeat the feeding and counter-feeding languages in Table 4-1.

It was found in the previous chapter that the observed empirical asymmetries were accounted for under the noisy-channel model by the interplay of two factors: ① the number of lexicons and grammars compatible with the data, and ② the number of lexicons and

grammars that were nearly compatible with the data. The first factor was found to be affected by the informativity of the OR, where different ORs either restrict or have no effect on the space of consistent hypotheses. I found that the resulting space roughly corresponded to the joint effect of both the Maximum Utilization bias and Transparency bias. The second factor was found to be affected by the relative frequency of the [t] versus [tʃ] alternants, which varied across each language, reminiscent of the Uniformity bias. For example, we see in Table 4-1 that the relative ratio of each respective alternant in the aforementioned alternation across all forms of the feeding and counter-feeding paradigms is 3:2 versus 4:1, respectively. This is due to the difference in the interacting OR for the alternating coronal-final stems in each language: the relevant interacting OR for the feeding language is [ikitʃi], while the relevant interacting OR for the counter-feeding language is [ikiti]. The presence of an additional [t] alternant raises the relative likelihood of hypotheses in which the palatalized consonant is lost on the surface, towards a state of greater uniformity. As such, depending on how many examples containing the non-palatalized coronal consonant are given to the model, we observe different results: under a distribution in which each slot in the paradigm is given uniformly, the feeding language is better replicated (Prickett 2019), whereas under a skewed distribution in which more examples of the stem in isolation are given, the counter-feeding ER is preferred (Kim 2012). In other words, related yet distinct surface patterns become less distinct as the relative ratio of alternants becomes increasingly skewed.

Crucially, ORs are capable of exhibiting the second factor independently from the first factor. For example, adding more ORs like the stem in isolation [ikit] increases the relative ratio of [t] versus [tʃ] tokens in the data, but does not provide any disambiguation on the phonological processes involved in the language, i.e. it is UNINFORMATIVE; it is compatible with the lexicon-grammar /ikit/ $\xrightarrow{\text{PAL.}}$ [ikit] as well as the lexicon-grammar /ikit/ $\xrightarrow{\varnothing}$ [ikit]. In this computational experiment, I provide an initial exploratory analysis probing deeper into the individual and joint effects of these two properties. Specifically, I aim to achieve two goals: ① to verify the claims asserted above, and ② to understand how these effects emerge as the relative distribution of not only the faithful ORs, but also the alternating ORs in the data are adjusted. I accomplish this task by manipulating the relative numbers of each alternant in the overall language, as well as the relative numbers of each OR type. These distributions are described in more detail in the following subsection.

4.1.2 Data manipulation and methods

In order to investigate the relative effects of each of the properties discussed above, I manipulated the relative frequences of certain forms observed in the alternating coronal-

final stem paradigms in the revised Baković languages. In particular, I manipulate the relative frequences of three different lexical contexts: ① the FAITHFUL contexts, e.g. [akita] $\langle \mu_1, \mu_B \rangle$, ② the PALATALIIZING contexts, e.g. [akitʃi] $\langle \mu_1, \mu_A \rangle$, and ③ the INTERACTING contexts, e.g. [akitʃi] or [akiti] $\langle \mu_1, \mu_A, \mu_B \rangle$ for the feeding and counter-feeding languages, and [akita] or [akitʃa] $\langle \mu_1, \mu_B, \mu_A \rangle$ for the bleeding and counter-bleeding languages. Depending on which lexical context is presented most frequently to the model, I generate one of four conditions: the ① UNIFORM distribution, ② FAITHFUL distribution, ③ PALATALIZING distribution, and ④ INTERACTING distribution.[1] Sample distributions for the feeding languages under the faithful, palatalizing, and interacting conditions are given in Table 4-2.

Under the uniform distribution, the model is provided with three complete paradigms consisting of all four forms and three incomplete paradigms consisting only of the stem in isolation. The remaining three conditions were generated from the uniform distribution by eliminating all but the target lexical contexts for two of the three complete paradigms. For example, the faithful distribution consists of a single complete paradigms and five incomplete paradigms. Two of the incomplete paradigms would be generated by taking two of the complete paradigms from the uniform distribution and eliminating all of the ORs for the paradigm except for the faithful lexical context, e.g. for the bleeding language, [ikata] $\langle \mu_2, \mu_B \rangle$ and [akata] $\langle \mu_3, \mu_B \rangle$. The remaining three incomplete paradigms correspond to the same incomplete paradigms as in the uniform distribution, in which only the stem in isolation is given. Both the palatalizing and interacting distributions were generated in the same manner, except that the palatalizating lexical contexts, e.g. [ikatʃi] $\langle \mu_2, \mu_A \rangle$ and [akatʃi] $\langle \mu_3, \mu_A \rangle$, and interacting lexical contexts, e.g. [ikata] $\langle \mu_2, \mu_B, \mu_A \rangle$ and [akata] $\langle \mu_3, \mu_B, \mu_A \rangle$, were kept in each respective condition.

I am interested in examining the effect of the data on the results of the noisy-channel model. To this end, I fix the values of all the parameters of the model. The parameterizations of each component are given in Table 4-3. There are a few aspects of the parameterization worth highlighting. First, note that I set our noisy-channel parameter $\lambda = 3$, which corresponds to a model with moderate noise. As I have indicated in the previous chapter, some level of noise is necessary in order to capture the empirical results, as it allows the effect of the nearly consistent hypotheses, which arise from the relative distribution of alternants, to emerge. Second, note that the lexicon is biased towards positing a single UR across all lexical contexts rather than memorizing context-specific URs. I will find in the next section that the parameterization of the lexicon does not appear to significantly affect the performance of the model.

[1] Note that I do not vary the relative frequencies of the alternants for the deleting alternation, e.g. [ikaki] $\langle \mu_4, \mu_A, \mu_B \rangle$ versus [ikaka] $\langle \mu_4, \mu_B \rangle$. I forgo investigating this alternation as there was no observed difference in performance across all four languages for this alternation.

Table 4-2: Sample distributions for the feeding language from Baković (2011).

Faithful Distribution

akit	$\langle \mu_1 \rangle$	akita	$\langle \mu_1, \mu_A \rangle$	akitʃi	$\langle \mu_1, \mu_B \rangle$	akitʃi	$\langle \mu_1, \mu_A, \mu_B \rangle$
	$\langle \mu_2 \rangle$	ikita	$\langle \mu_2, \mu_A \rangle$		$\langle \mu_2, \mu_B \rangle$		$\langle \mu_2, \mu_A, \mu_B \rangle$
	$\langle \mu_3 \rangle$	akata	$\langle \mu_3, \mu_A \rangle$		$\langle \mu_3, \mu_B \rangle$		$\langle \mu_3, \mu_A, \mu_B \rangle$
ikak	$\langle \mu_4 \rangle$	ikaka	$\langle \mu_4, \mu_A \rangle$	ikaki	$\langle \mu_4, \mu_B \rangle$	ikaki	$\langle \mu_4, \mu_A, \mu_B \rangle$
ikat	$\langle \mu_5 \rangle$	???	$\langle \mu_5, \mu_A \rangle$	???	$\langle \mu_5, \mu_B \rangle$	???	$\langle \mu_5, \mu_A, \mu_B \rangle$
ikit	$\langle \mu_6 \rangle$	???	$\langle \mu_6, \mu_A \rangle$	???	$\langle \mu_6, \mu_B \rangle$	???	$\langle \mu_6, \mu_A, \mu_B \rangle$
akat	$\langle \mu_7 \rangle$	???	$\langle \mu_7, \mu_A \rangle$	???	$\langle \mu_7, \mu_B \rangle$	???	$\langle \mu_7, \mu_A, \mu_B \rangle$
akik	$\langle \mu_8 \rangle$	???	$\langle \mu_8, \mu_A \rangle$	???	$\langle \mu_8, \mu_B \rangle$	???	$\langle \mu_8, \mu_A, \mu_B \rangle$

Palatalizing Distribution

akit	$\langle \mu_1 \rangle$	akita	$\langle \mu_1, \mu_A \rangle$	akitʃi	$\langle \mu_1, \mu_B \rangle$	akitʃi	$\langle \mu_1, \mu_A, \mu_B \rangle$
	$\langle \mu_2 \rangle$		$\langle \mu_2, \mu_A \rangle$	ikitʃi	$\langle \mu_2, \mu_B \rangle$		$\langle \mu_2, \mu_A, \mu_B \rangle$
	$\langle \mu_3 \rangle$		$\langle \mu_3, \mu_A \rangle$	akatʃi	$\langle \mu_3, \mu_B \rangle$		$\langle \mu_3, \mu_A, \mu_B \rangle$
ikak	$\langle \mu_4 \rangle$	ikaka	$\langle \mu_4, \mu_A \rangle$	ikaki	$\langle \mu_4, \mu_B \rangle$	ikakɪ	$\langle \mu_4, \mu_A, \mu_N \rangle$
ikat	$\langle \mu_5 \rangle$	???	$\langle \mu_5, \mu_A \rangle$	???	$\langle \mu_5, \mu_B \rangle$	???	$\langle \mu_5, \mu_A, \mu_B \rangle$
ikit	$\langle \mu_6 \rangle$	???	$\langle \mu_6, \mu_A \rangle$	???	$\langle \mu_6, \mu_B \rangle$	???	$\langle \mu_6, \mu_A, \mu_B \rangle$
akat	$\langle \mu_7 \rangle$	???	$\langle \mu_7, \mu_A \rangle$	???	$\langle \mu_7, \mu_B \rangle$	???	$\langle \mu_7, \mu_A, \mu_B \rangle$
akik	$\langle \mu_8 \rangle$	???	$\langle \mu_8, \mu_A \rangle$	???	$\langle \mu_8, \mu_B \rangle$	???	$\langle \mu_8, \mu_A, \mu_B \rangle$

Interacting Distribution

akit	$\langle \mu_1 \rangle$	akita	$\langle \mu_1, \mu_A \rangle$	akitʃi	$\langle \mu_1, \mu_B \rangle$	akitʃi	$\langle \mu_1, \mu_A, \mu_B \rangle$
	$\langle \mu_2 \rangle$		$\langle \mu_2, \mu_A \rangle$		$\langle \mu_2, \mu_B \rangle$	ikitʃi	$\langle \mu_2, \mu_A, \mu_B \rangle$
	$\langle \mu_3 \rangle$		$\langle \mu_3, \mu_A \rangle$		$\langle \mu_3, \mu_B \rangle$	akatʃi	$\langle \mu_3, \mu_A, \mu_B \rangle$
ikak	$\langle \mu_4 \rangle$	ikaka	$\langle \mu_4, \mu_A \rangle$	ikaki	$\langle \mu_4, \mu_B \rangle$	ikaki	$\langle \mu_4, \mu_A, \mu_B \rangle$
ikat	$\langle \mu_5 \rangle$	???	$\langle \mu_5, \mu_A \rangle$	???	$\langle \mu_5, \mu_B \rangle$	???	$\langle \mu_5, \mu_A, \mu_B \rangle$
ikit	$\langle \mu_6 \rangle$	???	$\langle \mu_6, \mu_A \rangle$	???	$\langle \mu_6, \mu_B \rangle$	???	$\langle \mu_6, \mu_A, \mu_B \rangle$
akat	$\langle \mu_7 \rangle$	???	$\langle \mu_7, \mu_A \rangle$	???	$\langle \mu_7, \mu_B \rangle$	???	$\langle \mu_7, \mu_A, \mu_B \rangle$
akik	$\langle \mu_8 \rangle$	???	$\langle \mu_8, \mu_A \rangle$	???	$\langle \mu_8, \mu_B \rangle$	???	$\langle \mu_8, \mu_A, \mu_B \rangle$

The hypothesis space over the rules is identical to the one used in Chapter 3, consisting of a vowel deletion process and two context-specific and context-free palatalization and depalatalization processes. I repeat the space of atomic rules in (1).

Table 4-3: Parameterization and summary of each parameter for Experiment 1.

Parameter	Setting	Effect on the Model
θ	0.5	Prefer shorter prototype URs.
α	0.75	Prefer to reuse the prototype UR.
ψ	5	Prefer to generate contextual URs similar to the prototype UR.
λ	3	Prefer ERs similar to the ORs, but allow for some variation.

(1) a. Vowel Deletion: V → ∅ / _V
b. Palatalization: t → tʃ / _i
c. Generalized Palatalization: t → tʃ
d. Depalatalization: tʃ → t / _{C, #}
e. Generalized Depalatalization: tʃ → t

Evaluation of the model followed the same structure as before, wherein I examine the relative probability assigned to the intended versus alternate ERs for the incomplete paradigms. Note that only the incomplete paradigms in which the stem in isolation is given, denoted by the grey-shaded cells, were used in evaluation. The incomplete paradigms in which the stem in isolation is not provided, denoted by the black-shaded cells, were omitted.

4.1.3 Computational results and discussion

The results of the model for the non-uniform distributions are given in Figure 4-1.[2] Note that no matter how the relative frequencies of each paradigm type are adjusted, the qualitative asymmetries remain the same as what we have observed in Chapter 3: the model performs better on the faithful trials when trained on the anti-maximally utilizing bleeding and counter-feeding interactions, better on the palatalizing trials when trained on the maximally utilizing feeding and counter-bleeding languages, and better on the interacting trials when trained on the transparent bleeding and feeding languages. Despite this, I point out several key observations that can be made with respect to the change in absolute frequencies between conditions.

I predicted that as the ratio of alternants is increasingly skewed towards [t], performance for the lexical contexts containing the unpalatalized consonant would increase, and performance for the lexical contexts containing the palatalized consonant would decrease. However, as the asymmetry is inverted, the expected boost in performance is likewise reversed. To verify this, I contrast the relative performance across languages between the

[2] The uniform distribution is omitted from our discussion as it produced a more or less identical outcome to what was observed in Chapter 3 under the multiple paradigm condition.

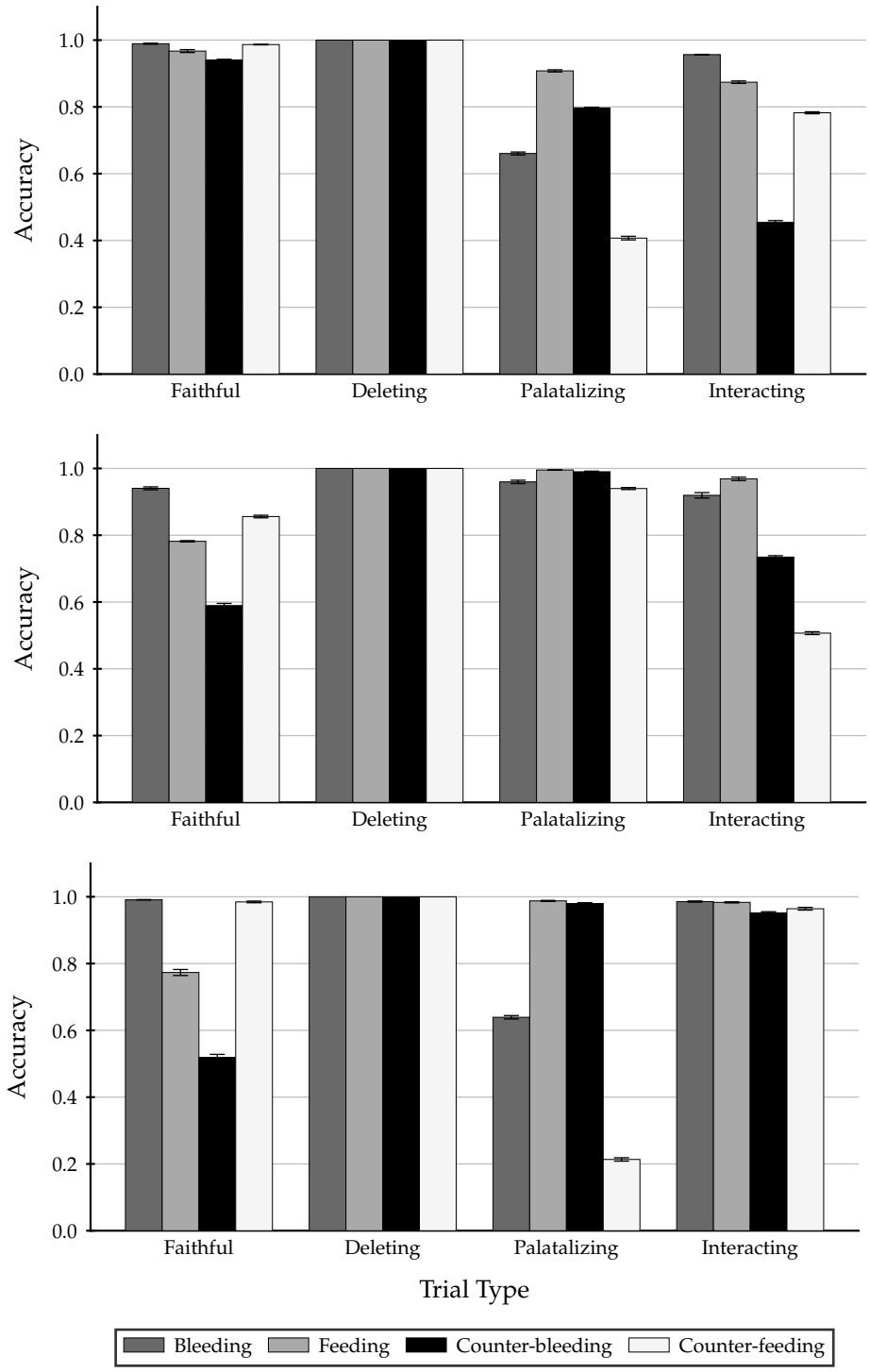
faithful and palatalizing conditions, given in the top and middle plots, respectively, in Figure 4-1. I will evaluate the claim by examining the respective differences in accuracy between the faithful and palatalizing distributions in the interacting trials.

We see that indeed, performance is drastically affected by the distribution of [t] versus [tʃ]. The model, when trained on the feeding and counter-bleeding languages, performs worse in the interacting trial given the faithful distribution than given the palatalizing distribution. In particular, note the difference in performance between the counter-bleeding and counter-feeding languages between the two conditions. In the faithful distribution, the model when trained on the counter-bleeding language performed extremely poorly, with it preferring the faithful, alternate ER *[ikata] over the opaque, intended ER [ikatʃa]; in contrast, the model when trained on the counter-feeding language performed relatively better, assigning most of the relative probability mass on the opaque, intended ER [ikiti] over the transparent, alternate ER *[ikitʃi]. This observation is reversed in the palatalizing condition, with the model trained on the counter-bleeding language performing above chance in the interacting trial, while the model trained on the counter-feeding language performing more or less at-chance. Despite the faithful form being minimally informative, as it does not provide any restrictions on compatible UR-grammar hypotheses, the performance of the model is affected nonetheless by mere presence of additional [t] forms. This verifies the claim that the number of [t] versus [tʃ] influences the kinds of hypotheses inferred by the model, independent of the informativity of the form itself.

How does the frequency of alternants interact with the informativity of the ORs given to the learner? I highlighted in the previous subsection that different ORs evoke different restrictions on the lexicophonological space. For instance, the faithful OR [ikata] does not provide any information regarding the existence or non-existence of either the palatalization or deletion processes, while, depending on the respective OR observed, the interacting OR assigns different levels of restrictivity on the hypothesis space. For example, the interacting OR for the counter-feeding language [ikati] can only be generated through the correct ordering of the palatalization and deletion processes given the UR /ikat-a-i/ or by underlearning palatalization, e.g. /ikat-_-i/ $\xrightarrow{\text{Pal.}}$ *[ikatʃi]. In contrast, the interacting OR for the feeding language [ikatʃi] has no such restriction, and is capable of positing hypotheses in which the OR is encoded underlyingly without restricting the space of compatible grammars associated with it, e.g. /ikatʃ-_-i/ $\xrightarrow{\text{Del.}}$ [ikatʃi] $\xrightarrow{\text{Pal.}}$ [ikatʃi]. To evaluate this, I compare the performance of the model between the faithful and interacting distribution, as well as the palatalizing and interacting distribution.

In the first comparison, I examine the performance of the model for the palatalizing trial given the faithful distribution versus the interacting distribution. Specifically, I examine the

Figure 4-1: Average accuracy of the model under different distributions, $\lambda = 3$. Top. The faithful distribution, Middle. The palatalizing distribution. Bottom. The interacting distribution.



difference in performance in the bleeding and counter-feeding languages. Distributionally, the languages both consist of the same ratio of [t] and [tʃ] alternants 8:1. If the number of each alternant is the only contributing factor shaping the results, we expect there to be no difference in the performance in the two languages between each distribution. However, if the informativity is also a contributing factor, we expect a there to be a difference in performance across the two distributions. We indeed see that this is the case: while the performance on the palatalizing trials for the counter-feeding language is relatively low in the faithful condition, it is even lower in the interacting condition, with the majority of the probability mass assigned to the non-palatalizing, alternate ER *[ikiti]. Interestingly, while there is a difference in performance in the bleeding language for the palatalizing trial, there is only a small dip in performance when comparing the accuracies under the faithful distribution versus the interacting distribution. This is indicative of how different the informativity of each interacting OR is for each language; while the counter-feeding OR restricts the space of acceptable hypothesis to only a handful of possible lexicon and grammar pairs, the bleeding OR only eliminates a few potential hypotheses, e.g. one in which palatalization precedes deletion /ikat-i-a/ $\xrightarrow{\text{Pal.}}$ [ikatʃia] $\xrightarrow{\text{Del.}}$ *[ikatʃa].

In the second comparison, I examine the performance of the model for the faithful trial given the palatalizing versus the interacting distribution. Specifically, I examine the difference in performance in the feeding and counter-bleeding languages. As above, these languages consist of the exact same ratio of [t] and [tʃ] 5:4. The difference lies in terms of what the value of the interacting OR is for each language. As before, if the number of alternants is the only predictor shaping the difference in acceptable hypotheses, we expect no difference in performance between the two conditions. If there is an effect of the informativity of the OR, we again anticipate there to be a difference in accuracies observed in the faithful trials for the feeding and counter-bleeding languages between distributions. As above, while we indeed to find a difference in performance, the difference is quite small. The reasoning behind it is the same: while the interacting OR does restrict the space of potential hypotheses, the relative restrictivity of the form is quite small.

The results of the exploratory analysis seems to indicate two initial conclusions. First, there is independent evidence to suggest that both the number of alternants as well as informativity of the individual forms play an active role in shaping the inferred space of hypotheses and thus the results. Second, while there is some effect of both properties on the weighted space, the contribution of informativity varies wildly between patterns. The weighted inferred lexicophonological space, then, primarily seems to be affected by the number of alternants in the data; only in certain cases such as the counter-feeding language do we observe a significant impact of informativity.

4.2 Experiment 2: How the Lexicon Shapes Generalization

In Chapter 2, I presented a noisy-channel model that jointly infers the URs and grammar given a space of unparsed ORs. A major component of this model consists of a robust generative model of the lexicon that is capable of positing non-phonological hypotheses over allomorphs, i.e. contextual allomorphy. This allows the model to encode surface alternations into the lexicon rather than or in addition to the phonology. In this section, I dive deeper into the role the lexicon plays in shaping the inferred lexicophonological space and consequently the predictions of the model. Specifically, I will demonstrate that the lexicon plays two separate roles in shaping the space of compatible hypotheses and nearly compatible hypotheses, respectively. I will show that the asymmetry in the replicability of the pattern is not affected at all by capacity of the model to posit unique, context-specific URs, which primarily dictates the space of compatible hypotheses. Rather, the performance of the model is almost exclusively handled by the phonology, with the lexicon only providing alternative starting points for different phonological strategies to converge on output distributions nearly consistent with the data. I will ultimately conclude from this that the observed asymmetry in model performance stems primarily from the distribution of nearly compatible grammars rather than the distribution of compatible grammars. The section is organized in the following manner. First, I provide the overview and motivation behind the experiment by disambiguating between the two ways in which the lexicon can shape the weighted space. Second, I introduce the manipulation over the parameterization of the lexicon that I will use in order to probe at which effects are actively playing a role in deriving the inferred space. Finally, I present the computational results and provide an exploratory commentary on the outcomes.

4.2.1 Overview and motivation

The lexicon is able to influence the distribution over consistent or nearly consistent lexicon-grammar hypotheses in two distinct ways. First, the generative lexicon allows surface alternations to either be encoded into the UR or derived from the phonology. Depending on the OR, different grammar hypotheses are compatible with a different number of UR hypotheses, resulting in a distribution that bolsters some grammar hypotheses over others. This largely correlates with the informativity of certain forms, as discussed in the previous section. Second, the generative lexicon expands the space of possible lexicon and grammar pairs to include many additional hypotheses that nearly capture the data. Depending on the distribution of possible outputs this space of lexicon-grammar pairs can now generate, the sheer number of hypotheses that are close enough to but not perfectly consistent with

the original set of ORs can overpower the set of consistent hypotheses. This is reminiscent of the effect the relative frequencies of alternants discussed in the previous section.

Each grammar is associated with a different space of compatible URs. These URs are possible due to the model's ability posit potentially-redundant contextual allomorphy. For example, given the interacting OR for the feeding language [ikatʃi] $\langle \mu_1, \mu_A, \mu_B \rangle$ and intended rule hypothesis consisting of deletion followed by palatalization, the learner can posit several distinct URs that produce the same output, such as those in (2).

(2) Compatible prototype and contextual UR hypotheses given the OR [ikatʃi] $\langle \mu_1, \mu_A, \mu_B \rangle$ and rule hypothesis ⟨Deletion, Palatalization⟩

a. /ikat-a-i/$_{\text{prototype}}$ → /ikat-a-i/$_{\text{contextual}}$ $\xrightarrow{\text{Del.}}$ [ikati] $\xrightarrow{\text{Pal.}}$ [ikatʃi]

b. /ikat-a-i/$_{\text{prototype}}$ → /ikat-_-i/$_{\text{contextual}}$ $\xrightarrow{\text{Del.}}$ [ikati] $\xrightarrow{\text{Pal.}}$ [ikatʃi]

c. /ikat-a-i/$_{\text{prototype}}$ → /ikatʃ-a-i/$_{\text{contextual}}$ $\xrightarrow{\text{Del.}}$ [ikatʃi] $\xrightarrow{\text{Pal.}}$ [ikatʃi]

d. /ikat-a-i/$_{\text{prototype}}$ → /ikatʃ-_-i/$_{\text{contextual}}$ $\xrightarrow{\text{Del.}}$ [ikatʃi] $\xrightarrow{\text{Pal.}}$ [ikatʃi]

Contrast this with the interacting OR for the counter-feeding language [ikati] $\langle \mu_1, \mu_A, \mu_B \rangle$. The intended underlying grammar of palatalization followed by deletion has comparably fewer compatible UR hypotheses; in fact, only one such UR hypothesis is compatible with this grammar. This is shown in (3).

(3) Compatible prototype and contextual UR hypotheses given the OR [ikati] $\langle \mu_1, \mu_A, \mu_B \rangle$ and rule hypothesis ⟨Palatalization, Deletion⟩

a. /ikat-a-i/$_{\text{prototype}}$ → /ikat-a-i/$_{\text{contextual}}$ $\xrightarrow{\text{Pal.}}$ [ikatai] $\xrightarrow{\text{Del.}}$ [ikati]

We see here that the lexicon allows for the learner to posit multiple different UR hypotheses for the same rule hypothesis, the space of which can vary depending on the OR. In terms of the inferred posterior distribution, then, relatively more probability can be in principle assigned to the intended feeding grammar than the counter-feeding grammar.

Note however that these alternative hypotheses come at the cost of having to posit a unique contextual UR for this lexical context. The question thus depends on whether the cost of positing a superfluous contextual UR is outweighed by the benefit of potentially having more hypotheses to assign posterior probability to.

In addition to the effect the lexicon has on the space of compatible hypotheses, I found that the lexicon is also capable of playing a significant role in dictating the structure of nearly consistent hypotheses. The basic story is the same as above: certain languages are more similar to some languages than others, and some of these languages have more compatible hypotheses associated with them. For example, the counter-feeding language

given in Table 4-1 is nearly compatible with a language in which no palatalized consonant emerges at all in the language. This language is consequently compatible with numerous lexicon and grammar pairs. For example, one hypothesis compatible with this alternative language is one in which no palatalization process is learned. This is demonstrated in (4).

(4) Underlearning the palatalization process for the counter-feeding language

Underlying Form	/ikit/	/ikit-a/	/ikit-i/	/ikit-a-i/
V → ∅ / _V	–	–	–	[ikiti]
Observed Form	[ikit]	[ikita]	[ikiti]	[ikiti]

Another possible hypothesis is one in which the UR of the alternating stem consists of the palatalized consonant /tʃ/ which is then depalatalized across all contexts, as shown in (5).

(5) Learning a generalized depalatalization process for the counter-feeding language

Underlying Form	/ikitʃ/	/ikitʃ-a/	/ikiʃt-i/	/ikitʃ-a-i/
tʃ → t	[ikit]	[ikita]	[ikiti]	[ikitai]
V → ∅ / _V	–	–	–	[ikiti]
Observed Form	[ikit]	[ikita]	[ikiti]	[ikiti]

Note that despite fact that the contextual underlying form for each lexeme is identical across all contexts, incorporating these hypotheses alters the space of grammars compatible with certain output distributions.

These strategies are not identical across all languages. For example, let us examine the feeding language. It is clear that the similarity between the feeding language and the alternative language presented above is not particularly high. However, this language is indeed similar to a different alternative language, one in which palatalization applies across all contexts except when found word-finally. Like with the alternate counter-feeding pattern, this pattern is also compatible with several different lexicon-grammar strategies. For example, one such hypothesis has the palatalized consonant /tʃ/ appear in the UR, which is then depalatalized in word-final position. This is demonstrated in (6).

(6) Learning a contextual depalatalization process for the feeding language

Underlying Form	/ikitʃ/	/ikitʃ-a/	/ikiʃt-i/	/ikitʃ-a-i/
tʃ → t / _#	[ikit]	–	–	–
V → ∅ / _V	–	–	–	[ikitʃi]
Observed Form	[ikit]	[ikitʃa]	[ikitʃi]	[ikitʃi]

As can be observed, each distribution is compatible with a different set of alternative patterns, which in turn are compatible with a different set of hypotheses.

In this subsection, I have demonstrated yet another source that can potentially shape the inferred lexicophonological space: the hypothesis space over the lexicon. I found two potential aspects of the lexicon that can alter this space: ① given that the model can posit unique URs for each lexical contexts, different ORs are associated with different numbers of compatible UR-rule pairs, thus creating an asymmetry in the space of compatible hypotheses, and ② lexical inference allows the model to consider alternative derivations that generate outputs nearly consistent with the data. I wish to determine whether each property contributes meaningfully to the predictions made by the model. However, the space is too large for us to individually enumerate and verify the number of consistent and nearly consistent hypotheses. To this end, I perform a computational experiment to investigate this problem. I outline the details of the experiment in the following subsection.

4.2.2 Parameter manipulation and methods

In order to investigate the effect of the lexicon on the performance of the model, I examined how the model's performance changed as I manipulated the parameter dictating the flexibility of the model to posit unique contextual URs: the identity hyperparameter α. This hyperparameter determines the rate at which the prototype UR of a lexeme is reused when positing a contextual UR for each lexical context in which a lexeme is found. Given a relatively lower parameterization, the model will be more willing to deviate from the prototype UR and propose a context-specific contextual UR. In contrast, given a higher parameterization, the model will be more willing to reuse the prototype UR, applying the same form across all contexts. I compare the model's performance over three values: $\alpha \in \langle 0.5, 0.75, 1 \rangle$.

I am interested in examining the effect of solely the lexicon on the results of the noisy-channel model. To this end, I fix the values of all other parameters of the model. The parameterizations of each component are given in Table 4-4. Note that this parameterization

Table 4-4: Parameterization and summary of each parameter for Experiment 2.

Parameter	Setting	Effect on the Model
θ	0.5	Prefer shorter prototype URs.
α	*varies*	Affects how often to reuse the prototype UR.
ψ	5	Prefer to generate contextual URs similar to the prototype UR.
λ	3	Prefer ERs similar to the ORs, but allow for some variation.

is identical to the one used in the previous experiment, except that here I vary the value of α. Furthermore, I wish to highlight that while the ψ, which dictates the level of variation from the prototype UR, is set rather high in the parameterization, I found that it was necessary in order for the model to successfully converge. Independent testing of the parameterization on the toy language presented in Chapter 2 appears to indicate that this parameter does not significantly affect model performance.

4.2.3 Computational results and discussion

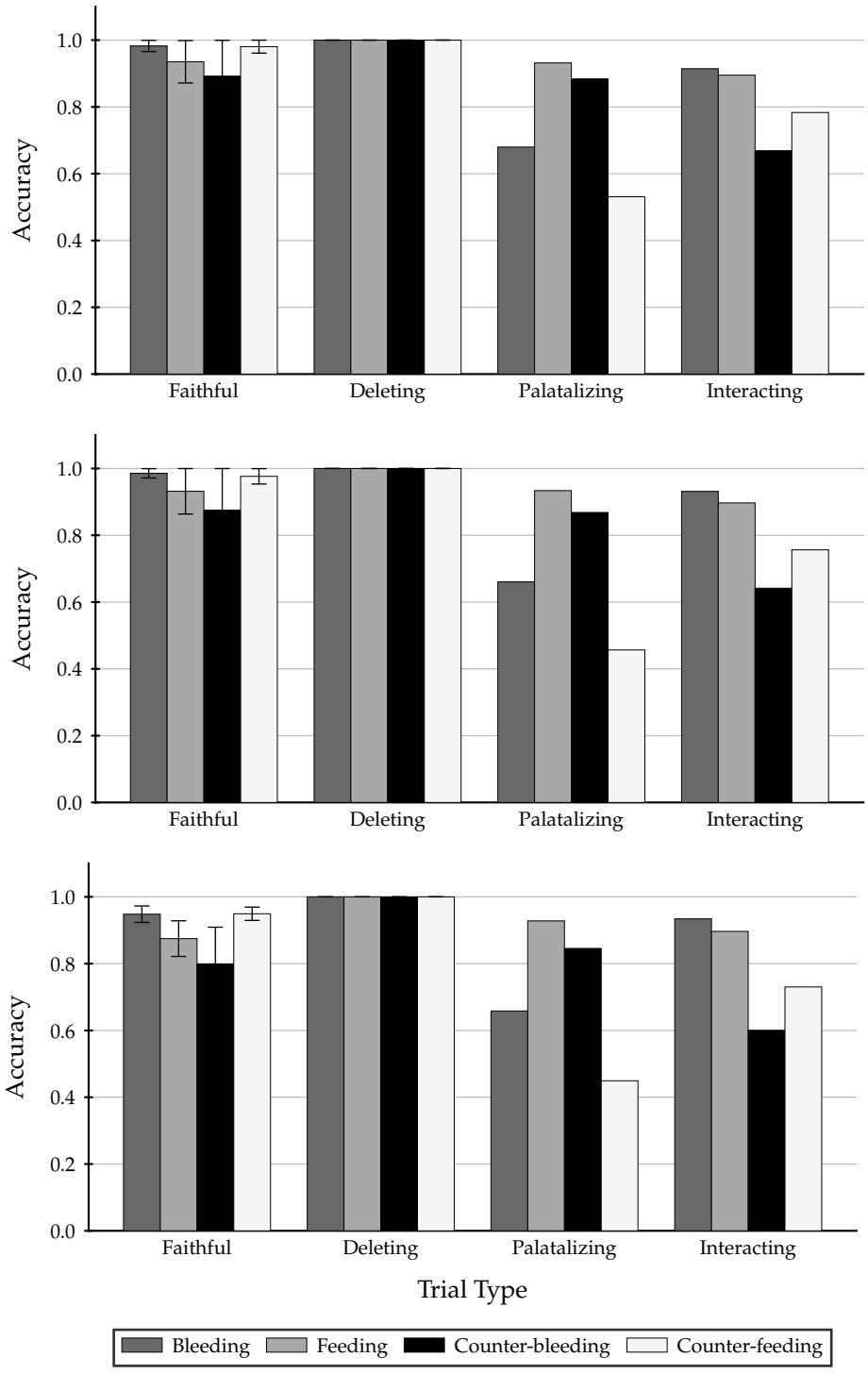
The results of the model are given in Figure 4-2. If the model is significantly affected by the lexicon's ability to posit context-specific URs, we expect there to be some variation in the performance of the model as the main parameter dictating this behavior, α, is adjusted. We see however that as we change the value of α, the results are almost completely unaffected. The only noticeable difference lies in the faithful trials, as performance minimally decreases as the value of α increases.

This seems to suggest that despite allowing the lexicon the ability to posit distinct contextual URs, the model heavily disprefers this strategy. Instead, it appears that the lexicon is playing only a very specific role: it allows the grammar flexibility in what phonological analyses are proposed and accepted. With a high enough level of noise, the model is allowed to consider drastically different phonological analyses that nearly capture the data, not only by way of underlearning a phenomenon, e.g. underlearning palatalization for the counter-feeding language, but also by learning a completely different grammar entirely, e.g. positing an underlying /tʃ/ with a word-final depalataliztion for the feeding language.

4.3 Relation to Human Learning

In this chapter, I investigated the relationship between different aspects of the data and the lexicon in shaping the lexicophonological space, and consequently the predictions and

Figure 4-2: Average accuracy of the model under different parameterizations over the lexicon, $\lambda = 3$. From top to bottom, $\alpha \in \langle 0.5, 0.75, 1 \rangle$.



performance of the model. I found that while there are two aspects of the data that could potentially influence the inferred weighted space of hypotheses, that of the informativity of the data as well as the number of alternants in the data, the latter effect was the predominant force in shaping the results under a noisy parameterization. Likewise, while there are two aspects of the lexicon that could potentially influence the inferred weighted space, that of the number of associated UR hypotheses associated with a grammar hypothesis compatible with the data as well as the number of associated UR hypotheses associated with a grammar hypotheses nearly compatible with the data, I found again that the latter property was the primary – if not only – component that determined the performance of the model. Interestingly, the results of both experiments converges on the same conclusion: the space of nearly consistent hypotheses serves as the driving force behind determining the predictions of the model under a noisy parameterization. This is in line with observations made by Rafferty and colleagues (2013). They suggested that the persistence of a pattern, and thus its typological frequency, is determined from the intersection of three properties: ① the learnability of the pattern, ② the distributions of different surface patterns, and ③ the similarity of these patterns to the observed distribution. All three properties are encompassed in our noisy-channel model; translated into our model's terms, the persistence of a pattern is determined by the aggregated weighted posterior over possible surface distributions given the data. Despite some patterns providing more restrictive data to the model, i.e. the counter-feeding OR [ikiti] can only be generated from a very small space of compatible hypotheses and consequently the intended underlying hypotheses have a higher relative posterior than the intended underlying hypotheses for the other languages, replicability is determined primarily by how similar the surface pattern is to other patterns.

Before concluding the chapter, I wish to address a key question that arises under our use of this model: what is the relationship between this model and human language acquisition? The results of the model are indicative of POPULATION-LEVEL performance; that is, it indicates the volatility of a language assuming a population of speakers, each contributing some amount of probability to the posterior. However, how do we connect this population-level model to individual human performance?

One possibility is that every learner of the language has full access to the entire distribution over the inferred weighted space, contrary to the more common assumption that learners converge on only a single grammar. This was suggested under the Multiple Grammars Theory of Variation (Kiparsky 1965; Anttila et al. 2007). Under this hypothesis, a learner samples from the entirety of the weighted space every time they generate a word is produced. However, this does not seem to be an adequate explanation for my model here. The Multiple Grammars Theory of Variation predicts that when aggregating over

all production tokens for an individual speaker, these productions would follow the same distribution as what was seen here. We have seen evidence in a multitude of areas, however, that individual speakers make unique and distinct predictions in certain contexts, even in the experimental domain (Ettlinger, Bradlow, & Wong 2014; Martin & White 2021). We thus need a more nuanced response.

I suggest one possible way of translating this model into individual performance: learners at the end of inference arrive over a truncated distribution over possible lexicons and grammars, sampled from the posterior. In other words, once the posterior distribution is estimated, learners will sample a subset of the hypotheses based on their relative posteriors. Thus, each individual may have slightly different individual hypotheses at the end of learning, but over the entire population, the empirical observation is retained.

Chapter 5

Reflecting on the Lexicophonological Space, and Areas of Future Research

In this chapter, I connect the results of the noisy-channel lexicophonological learner back to the original goals of the dissertation as well as to the existing literature, and introduce areas of future research. The chapter is organized into the following three sections. ① I summarize the main findings of the computational experiments performed in Chapters 3 and 4, and reflect on the contributions they make to the original questions posed at the beginning of the dissertation. I compare the results to conclusions made in previous work, and ultimately argue that the results of the model point back to the same basic claim: some surface sound patterns are more difficult to generalize because the data is too ambiguous to recover the original underlying grammar used to generate the pattern, resulting in mislearning. The ambiguity of the data is modulated by two factors: the informativity of each individual form, as well as the relative frequency of forms in the data. ② I discuss how changing our assumptions about the hypothesis space over the grammar can potentially change the outcome of the model. I illustrate how assigning a prior over rules, as proposed in previous research, can shift the weighted inferred lexicophonological space to favor attested but under-utilized alternative approaches such as contextual allomorphy. I also demonstrate that different theories of phonology, such as Optimality Theory, produce very different hypothesis spaces over the grammar, which in turn may result in different predictions of the model. ③ I provide a few avenues for future research, such as investigating the learnability of other kinds of process interactions like self-destructive feeding or seeding, as well as exploring the effect that surface distributions generated by process interactions on environment versus on focus may shape learnability.

5.1 The Lexicophonological Space and Learnability

The goal of the dissertation is to examine how the lexicon contributes to the replicability and generalizability of a surface sound pattern, using the learnability of process interactions as a case study. I demonstrated that the lexicon not only introduces additional alternative analyses to the data such as contextual allomorphy, but also vastly expands the space of grammatical hypotheses available to the learner. I proposed that the generalizability of a surface pattern emerges as a consequence of the relative number of hypotheses that are consistent with or nearly consistent with that pattern, and developed a novel noisy-channel model of lexicophonological learning in order to investigate this proposal.

In Chapter 3, I applied this model to two artificial language learning experiments that seemingly produced contradictory results, despite using the same underlying processes to generate the data. Prickett (2019) found that participants were better able to replicate and generalize forms generated by a feeding interaction than forms generated by a counter-feeding interaction. In contrast, Kim (2012) found that participants, in the absence of overt evidence of how the two processes interacted with each other, preferred to produce the counter-feeding form over the feeding form. The model is able to capture both results, as a result of two factors. First, it was found that certain process interactions produce surface distributions that are compatible with more lexicons and grammars than others, improving the ability of the model to generalize that pattern. However, the contribution of the space of consistent hypotheses is potentially offset by the contribution of the space of nearly consistent hypotheses. I found that depending on the relative frequencies of each form given to the model, the inferred hypothesis space may shift more of the probability mass towards hypotheses that generate nearly compatible, but ultimately distinct, alternative surface patterns instead. This results in the model performing better at generalizing certain forms while performing worse at generalizing others.

In Chapter 4, I examined in more detail how these factors interact with each other by performing two exploratory computational experiments and uncovered two important properties of the model. ① The observed asymmetries predicted by the model almost exclusively emerge due to the inferred space of nearly consistent hypotheses rather than the space of consistent hypotheses. ② Despite allowing the model to encode surface alternations as part of the lexicon rather than the grammar, the weighted inferred space contains extremely few hypotheses in which this option is used. In other words, the model is better able to generalize patterns generated by certain process interactions over others because they are compatible with more hypotheses in which the grammar retains the crucial rule ordering.

5.1.1 Reflecting on the current state of affairs

While the notions of the Maximal Utilization and Transparency learnability biases were first proposed as descriptive explanations for why certain process interactions were more likely to undergo historical change than others, computational modeling work from the outset strove to replicate the effects of these biases as an indirect consequence of the learning process. The underlying source driving the asymmetries in these models is grammatical ambiguity; some distributions are ambiguous between multiple grammatical hypotheses, which can only be distinguished by a relatively smaller set of forms in the data than others. As shown above, the noisy-channel model reaches an identical conclusion, providing further support for this hypothesis. However, the model offers additional context and contributions in a few key areas.

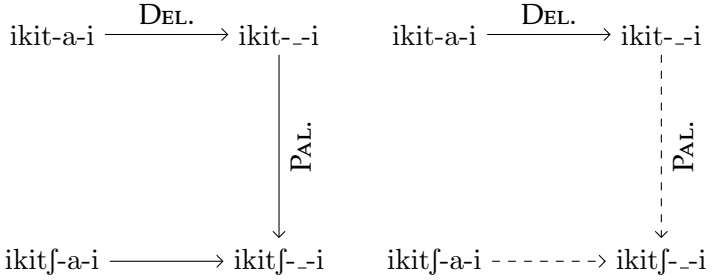
Unlike many past approaches exploring this topic, the noisy-channel lexicophonological model performs joint learning of the lexicon and grammar. Not only that, the model is also given the ability to posit contextual allomorphy, in which alternations are encoded into the UR directly as opposed to generating these alternations through the grammar. An interesting property of including UR learning is that it potentially introduces another form of ambiguity: LEXICAL AMBIGUITY. I demonstrated in Chapter 1 that the lexicon can influence the lexicophonological space differently depending on the distribution of the data. For example, the bleeding and counter-feeding languages incur a lower cost to hypotheses in which the alternation is lexicalized than the feeding and counter-bleeding languages do; the model needs to posit only one context-specific contextual UR for the former two languages, whereas it needs to posit two for the latter two languages. We saw in Chapter 4, however, that even if we prevent the model from utilizing contextual allomorphy by adjusting the relevant parameter $\alpha = 0$, the model still achieves the exact same qualitiative asymmetries, with minimal effect on performance. This suggests that the learnability asymmetry observed in prior computational work may indeed arise solely as a consequence of grammatical ambiguity, with little effect of lexical ambiguity.

In addition, previous computational models were successfully able to derive these effects as a result of how they performed search over the hypothesis space of grammars; either the model takes too long to learn the correct generalization and ends up in an incomplete state by the time learning ends (Jarosz 2016; Prickett 2019), or the model is more likely to converge on an incorrect grammatical analysis as a consequence of local optimization (Nazarov & Pater 2017). These results thus depend on the algorithmic implementation of search. As the noisy-channel model generates predictions over the entire inferred weighted space, the effect of grammatical ambiguity emerges instead as a formal characteristic of noise. Thus, the asymmetry predicted by the model emerges as a result

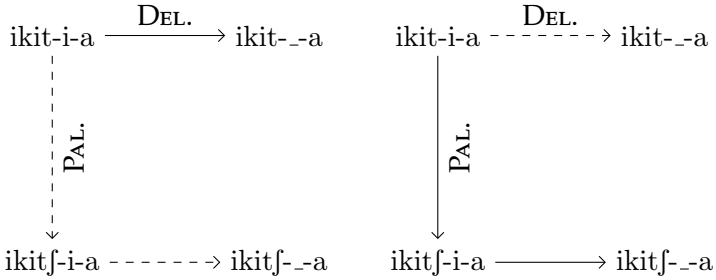
of LINGUISTIC COMPETENCE as opposed to LINGUISTIC PERFORMANCE (Chomsky 1965), and formalizes grammatical ambiguity by quantifying the relationship between the data and the space of possible lexicon and grammar hypotheses.

In Chapter 3, I noted an interesting observation while examining the space of compatible joint lexicon-grammar hypotheses: the volume of compatible hypotheses correlates with the asymmetrically-weighted joint effect of the Transparency and Maximal Utilization biases. Under these terms, it appears as though the Maximal Utilization bias affects this distribution more than the Transparency bias does. When examining the space of compatible lexicons and grammars, it is apparent why: while the Maximal Utilization bias emerges faithfully from this space, the lexicophonological space only encodes the negative effect of being not surface-true, with no differentiation between surface-apparent and not surface-apparent interactions. In order to aid in the discussion, I will once again use the formalism of input and output provision and removal. I repeat the respective graphs for the counter-feeding and feeding interactions, as well as bleeding and counter-bleeding interactions for the revised Baković (2011) languages in (1) and (2).

(1) Space of compatible joint lexicons and grammars for the feeding OR [ikitʃi] (LEFT) and counter-feeding OR [ikiti] (RIGHT)



(2) Space of compatible joint lexicons and grammars for the bleeding OR [ikita] (LEFT) and counter-bleeding OR [ikitʃa] (RIGHT)



The emergence of the Maximal Utilization bias in the lexicophonological space is straight-forward: if an interaction involves the application of both processes, the inter-

mediate form becomes an available UR hypothesis. In contrast, the supposed effect of the Transparency bias is detected by the kinds of edges leaving each respective OR. We see that all four interactions have a solid edge entering each respective interacting form, however, only in the case of the counter-feeding OR [ikiti] do we see a dashed edge leaving the form, resulting from deletion input-providing palatalization. The consequence of this was cast in terms of informativity: the intended phonological generalization cannot be applied with this UR, resulting in a more restrictive space and smaller respective distribution of consistent hypotheses. This seems to suggest that the model treats the properties of surface-apparentness and surface-trueness differently.

Furthermore, despite the model indirectly encoding aspects of both the Maximal Utilization and Transparency biases within the space of consistent hypotheses, I demonstrated that the majority of the observed asymmetries stems from the space of nearly consistent hypotheses. This space is dictated in large part by the distribution of the data, specifically the relative frequencies of each alternant in an alternation and the resultant change in the relative similarity of the given distribution to an alternative distribution. It has been demonstrated in past work that the relative type frequency of forms in the data shapes the learnability and productivity of a process (Bybee 2003; Yang 2005; Gerken & Bollt 2008). This general idea has recently been shown to also shape the relative learnability of process interactions (Jarosz 2016). The noisy-channel model provides support for this hypothesis. Process interactions generate distributions that may make it easier or harder for the learner to recover the original underlying grammar. However, depending on the relative frequencies of each form in the data, the relative learnability of a process interaction may change. This was demonstrated in the computational experiments explored in Chapter 3, where a learnability preference for the feeding or counter-feeding language emerges directly from the relative frequencies of each alternating form as a consequence of how similar the counter-feeding language is to one in which palatalization is not learned at all. The effect here is reminiscent of a Uniformity bias, in which a pattern is increasingly pushed towards a state of minimal contrast – here, the loss of the palatalization contrast. Previous theoretical work derived this effect by proposing a mechanism that pressures the grammar to keep certain correspondence between related forms, oftentime paradigms (Kenstowicz 1995; Benua 1997; Kager 1999, among others). The noisy-channel model derives the effect slightly differently, allowing the pressure of minimal contrast to surface regardless of paradigm assignment.

Lastly, the noisy-channel model makes different assumptions about the end-state of learning. Whereas previous approaches assume that the learner converges on a single, potentially probabilistic, grammar, the noisy-channel model assumes that the learner con-

verges on a probabilisic distribution over deterministic lexicon-grammar hypotheses. The model thus extends and supports the claims first made by Bane and Riggle (2008), as well as Rafferty et al. (2013) with respect to the learnability of process interactions.

5.2 The Grammatical Space and Learnability

In this section, I explore how adjusting our assumptions over the space of grammars can fundamentally change the outcome of the model. The section is organized into the following two subsections. ① I discuss the potential impact of adding an informative prior on the grammar to the predictions of the model. ② I illustrate how different theories of phonology are associated with vastly different hypothesis spaces over possible grammars. I focus on two specific proposals, Optimality Theory and Stratal Optimality Theory, in order to convey this point.

5.2.1 Exploring the impact of informative priors on the grammar

In Chapter 1, I reported that certain process interactions are more likely to be reanalyzed than others. I provided a non-exhaustive list of possible outcomes of reanalysis, such as contextual allomorphy, in which an alternation is lexicalized rather than learned as a productive phonological phenomenon. However, despite designing a model that includes this analysis in the space of lexical hypotheses, I found that the strategy is rarely, if at all, used. In other words, despite contextual allomorphy being well-attested in the literature, the noisy-channel model never assigns significant probability to these kinds of hypotheses. The reason for this is due to the assumptions made regarding the grammar; the model assumes a uniform prior over the grammar: a hypothesis consisting of multiple, specific rules is just as probable as a hypothesis consisting a fewer, more general rules.

The question of process induction has been a long-standing question of interest in rule-based theories, both in the theoretical (Halle 1962; Chomsky & Halle 1965, 1968) as well as the computational realm (Gildea & Jurafsky 1996; Peperkamp et al. 2006; Goodman et al. 2008; Calamaro & Jarosz 2015). Much research has converged on the proposal that the inferred grammar be maximally "simple." The notion of simplicity oftentimes incorporates one or more of the following conditions: ① the rules are minimally complex, and ② the number of rules is minimized. One way in which this preference is encoded is through a generative model over rules (Goldwater & Johnson 2004; Rasin et al. 2015; Ellis et al. 2022). The consequence of having a generative model over the phonology is clear: instead of allowing the grammar to freely posit any rule hypothesis with equal prior probability, proposing a rule comes at the cost of a lower prior. In the face of a surface alternation, then,

the learner must now weigh between learning a rule or lexicalizing the alternation. I leave investigating the effect the generative model has on the lexicophonological space to future research.

5.2.2 How different theories of phonology influence the grammatical space

Throughout this dissertation, I adopted the rule-based framework of SPE in order to evaluate the joint effect of the lexicon and grammar on learnability. This formulation was chosen as there is no innate representational bias towards certain process interactions over others.

In this subsection, I schematically demonstrate how altering our assumptions regarding the phonological theory changes the hypothesis space over grammars, and in turn the potential predictions of the model. The subsection is organized as follows. I first introduce and formally define an alternative class of theories that have been proposed to capture phonological phenomena: constraint-based theories, such as Optimality Theory (Smolensky & Prince 1993). I discuss how this system represents or fails to represent different process interactions, and how this in turn may affect the inferred weighted lexicophonological space. I then introduce an augmentation of the framework, Stratal Optimality Theory (Kiparsky 2000), and discuss the differences and similarities of the resultant inferred weighted space to the previously mentioned proposals.

Optimality Theory

Optimality Theory, henceforth OT, is a constraint-based theory of phonology that generates ERs based on the elimination of candidates via ranked violable constraints. The formalism adopts three core components: ① Gen, or the function that takes in an underlying form and returns the set of potential candidate forms, ② Con, the distribution of constraints to which candidates will be assessed and eliminated, and ③ Eval, or the function that takes in the underlying form and candidate pair, as well as the constraint violations of each pair, and returns a single or a set of remaining candidates. I focus our attention on the latter two components of Con and Eval for the discussion here.

Con in its strictest interpretation comprises the universal set of violable constraints that all languages are derived from. These constraints typically fall under two large classes. The first class corresponds to the markedness constraints, which assign violations to surface configurations. To understand how markedness constraints work, take the familiar rules given in (3) as an example.

(3) a. Deletion: $V \rightarrow \varnothing$ / _V
b. Palatalization: $t \rightarrow t\int$ / _i

These rules correspond to a structural description, as well as a transformation that takes place when that structural description is met. We can restate the structural description in terms of a surface restriction: a transformation occurs because the context for which a sound is found in is dispreferred on the grounds of difficulties in production or perceptibility. The structural descriptions of the deletion and palatalization processes discussed throughout this dissertation can therefore by restated into the markedness constraints in (4).

(4) *VV Assign a * for every instance in which a vowel precedes another vowel.
*ti Assign a * for every instance the segment [t] precedes the segment [i].

The second class corresponds to the FAITHFULNESS CONSTRAINTS, which assign violations to changes from the input to an output. For example, the application of palatalization to the input /ikiti/ $\xrightarrow{\text{PAL.}}$ [ikitʃi] incurs one change: the change from [−dist] to [+dist]. We can define the transformation denoted in the deletion and palatalization rules as in (5).

(5) MAX-IO Assign a * when an input segment has no output correspondent.
IDENT-IO-[dist] Assign a * when two corresponding segments differ in [dist].

Constraints are ranked in order of priority of elimination: violations of higher-ranked constraints are dispreferred relative to violations of lower ranked constraints. EVAL corresponds to the function that, given a sequence of ranked constraints and potential candidates for a given input, determines the output by sequentially eliminating candidates until either a single candidate remains or all the constraints have been examined. For example, given the ranking in (6) and input /ikit-a-i/, EVAL predicts the optimal candidate to be [ikitʃi].

(6) *ti, *VV » MAX-IO » IDENT-IO-[dist]

/ikit-a-i/	*ti	*VV	MAX-IO	IDENT-IO-[dist]
[ikitai]		*!		
[ikiti]	*!		*	
☞ [ikitʃi]			*	*

Crucially, OT is not capable of natively representing opaque interactions as a phonological phenomenon. Embedded in all of the OT variants is the notion of SURFACE EVALUATION: outputs are generated through the evaluation of the surface form, not on the intermediate representations. Thus, forms in which the intended winning output violates a surface generalization observed elsewhere in the language, as is the case for underapplication, will lose to the candidate in which the generalization is satisfied. For example, let us specify

the rankings needed to capture the revised Baković (2011) counter-feeding language.

In order to capture the deletion process /ikik-a-i/ $\xrightarrow{\text{Del.}}$ [ikitʃi], *VV must outrank Max-IO, as demonstrated in (7).

(7) *VV » Max-IO

/ikik-a-i/	*ti	*VV	Max-IO	Ident-IO-[dist]
[ikikai]		*!		
☞ [ikiki]			*	

In order to capture the palatalization process /ikit-i/ $\xrightarrow{\text{Pal.}}$ [ikitʃi], *ti must outrank Ident-IO-[dist]. Moreover, in over to favor palatalizing the consonant as opposed to deleting it, either *VV or Max-IO must outrank Ident-IO-[dist]. This is demonstrated in (8).

(8) *ti, {*VV, Max-IO} » Ident-IO-[dist]

/ikit-i/	*ti	*VV	Max-IO	Ident-IO-[dist]
[ikiti]		*!		
[ikit]			*!	
☞ [ikitʃi]				*

Now consider the interacting OR for the counter-feeding language [ikiti]. The relevant tableau is given in (9).

(9) Underapplication cannot be represented in OT

/ikit-a-i/	*ti	*VV	Max-IO	Ident-IO-[dist]
[ikitai]		*!		
☛ [ikiti]	*!		*	
☞ [ikitʃi]			*	*

As can be observed, given the set of constraints and necessary rankings outlined above, the intended winning candidate [ikiti] loses to the alternative candidate [ikitʃi]; the constraint *ti is violated by the intended candidate, but, as there is no higher-ranking constraint to compel this violation, it loses to the observed alternative candidate.

In addition to OT being unable to represent cases of underapplication, the theory more-

over fails to phonologically capture outputs that are generated through the gratuitous application of a process without explicit need on the surface, as is the case of overapplication. For example, let us specify the necessary rankings needed to capture the revised counter-bleeding language from Baković (2011). This language is trivially identical in all respects to the counter-feeding language except in the interacting lexical context, in which the respective OR is [ikitʃa]. The relevant tableau is given in (10).

(10) Overapplication cannot be represented in OT

/ikit-i-a/	*ti	*VV	MAX-IO	IDENT-IO-[dist]
[ikitia]	*!	*		
☞ [ikita]			*	
☛ [ikitʃa]			*	!*

As can be observed, the palatalization of the coronal segment [t] is not justified by any additional improvement in markedness; the palatalization process is gratuitous.

There are two consequences that emerge from moving from rule-based formalisms to constraint-based theories such as OT. First, note that the necessary ranking conditions that must be satisfied do not specify the rankings of every constraint. This results in PARTIAL RANKINGS, in which unranked constraints can freely vary between either rankings (Anttila et al. 2007; Jarosz 2006b). As such, languages that require fewer rankings to define are consequently compatible with a larger number of complete rankings. This was an observation made by Riggle (2010) and indirectly stated by Stanton (2016), who argued that certain patterns were easier to learn under the Gradual Learning Algorithm model (Boersma et al. 1997; Boersma & Hayes 2001) as they required a subset of the necessary nested rankings to reach a consistent outcome. Second, the noisy-channel model adopting an OT framework would be biased in favor of the transparent interactions, as the only way of achieving the opaque counter-bleeding and counter-feeding interactions would have to be through lexicalization.

Many amendments to the theory have adjusted different aspects of the base OT model in order to accommodate opacity. These include adjusting EVAL by incorporating serial derivation, such Stratal OT (Kiparsky 2000), or by modifying CON and deriving opaque interactions from additional independently-motivated constraints such as Output-Output constraints (Kenstowicz 1995), or by incorporating special constraints to account for the opaque interaction, such as Sympathy Theory (McCarthy 1999) Conjoined Constraints (Smolensky 1995; Ito & Mester 2003), OT with Candidate Chains (McCarthy 2007), and

Serial Markedness Reduction (Jarosz 2014). Each of these modifications result in distinct hypothesis spaces over grammars, which in turn may result in different predictions of the model. Crucially, however, despite now allowing all process interactions to be encoded as purely phonological phenomena, most of the grammatical spaces under each respective proposal, unlike rule-based formalizations, contain an indirect bias in favor of transparent interactions. Due to space considerations, I examine the lexicophonological space that emerges under the Stratal OT approach and leave consideration of the alternative OT models to future research.

Stratal Optimality Theory

Unlike in classical OT, in which evaluation is done within a single evaluative cycle, Stratal OT assumes that evaluation occurs over multiple strata. Each stratum moreover to have its unique constraint ranking, with the optimal candidate from the initial stratum then used as the input to the next stratum. Both the counter-bleeding and counter-feeding patterns require at least two strata in order for the phenomenon to be represented, but the requirements for each vary drastically. In order to capture the counter-feeding OR [ikiti], the palatalization and deletion processes must be encoded in different strata. In the first stratum, palatalization must occur to the exclusion of deletion. This is accomplished with the ranking MAX-IO » *VV, *ti » IDENT-IO-[dist]. In the second stratum, deletion may now apply, however, this time to the exclusion of palatalization. This is accomplished with the ranking *VV » MAX-IO, IDENT-IO-[dist] » *ti. In contrast, for the counter-bleeding OR [ikitʃa], while the palatalization and deletion processes must be encoded into different respective strata, the necessary ranking conditions within each stratum are slightly different. In the first stratum, like with the counter-feeding language, palatalization must occur to the exclusion of deletion. In the second stratum, however, so long as deletion applies, encoded via the ranking *VV » MAX-IO, the respective rankings of the other constraints do not matter. The respective tableaux for each opaque interaction are given in (11) and (12).

(11) Representing counter-feeding in Stratal OT
STRATUM 1: MAX-IO » *VV, *ti » IDENT-IO-[dist]

/ikit-a-i/	MAX-IO	*VV	*ti	IDENT-IO-[dist]
☞ [ikitai]		*		
[ikiti]	*!		*	
[ikitʃi]	*!			*

Stratum 2: *VV » Max-IO, Ident-IO-[dist] » *ti

	/ikit-a-i/	*VV	Max-IO	Ident-IO-[dist]	*ti
	[ikitai]	*!			
☞	[ikiti]		*		*
	[ikitʃi]		*	*!	

(12) Representing counter-bleeding in Stratal OT

Stratum 1: Max-IO » *VV, *ti » Ident-IO-[dist]

	/ikit-i-a/	Max-IO	*VV	*ti	Ident-IO-[dist]
	[ikitia]		*	*!	
☞	[ikitʃia]		*		*
	[ikita]	*!		*	
	[ikitʃa]	*!			*

Stratum 2: *VV » Max-IO, Ident-IO-[dist] » *ti

	/ikitʃ-i-a/	*VV	Max-IO	Ident-IO-[dist]	*ti
	[ikitia]	*!		*	*
	[ikitʃia]	*!			
	[ikita]		*	*!	
☞	[ikitʃa]		*		

We observe that different interactions under Stratal OT have inherent differences in the grammatical space: the bleeding and feeding interactions posit the exact same ranking conditions, and can accommodate the entire language through a single stratum. In contrast, the opaque interactions require the processes to be split across two strata, but the minimum requirements differ between the two, with the counter-feeding language requiring more ranking conditions than the counter-bleeding language. This results in the counter-feeding language having a smaller space of compatible grammars relative to the counter-bleeding language. As such, examining just the space of compatible grammars given the respective URs /ikit-a-i/ and /ikit-i-a/ for the feeding and counter-feeding, as well as the bleeding and counter-bleeding languages, respectively, we observe that the number of compatible

grammars in order from highest to lowest corresponds roughly to what was seen in the interacting trials in Prickett (2019): feeding and bleeding at the top, followed by counter-bleeding, and finally counter-feeding. It is apparent that changing the theory utilized by the model results in a different space of grammars. I again leave exploring the impact of this space on the results of the model to future research.

5.3 Additional Properties of Process Interactions and Learnability

In this section, I explore a few areas of potential future research examining the effect certain properties of process interactions may have on learnability. I explicate two such properties: ① the relationship between the process interaction and its respective lexicophonological space, and ② the formal notions of interactions on environment versus on focus. I briefly discuss the consequences of each in their respective subsections below.

5.3.1 How does a process interaction relate to its lexicophonological space?

I have assumed the relationship between the process interaction and its respective lexicophonological spaces to be a primitive. For example, feeding interactions are compatible with more lexicons and grammars than counter-bleeding interactions, and counter-bleeding interactions are compatible with more lexicons and grammars than bleeding interactions. However, there exist several cases in which this correspondence is not accurate. I focus on a particular example, SEEDING, for our discussion here.

Given two processes A and B, seeding (Baković & Blumenfeld 2018, 2022) is an over-application phenomena in which A feeds B, but the conditioning environment for A is eliminated from the surface as a result of B's application. A canonical example of this is that of i-epenthesis and k-deletion in Turkish, outlined in rule-based terms in (13).

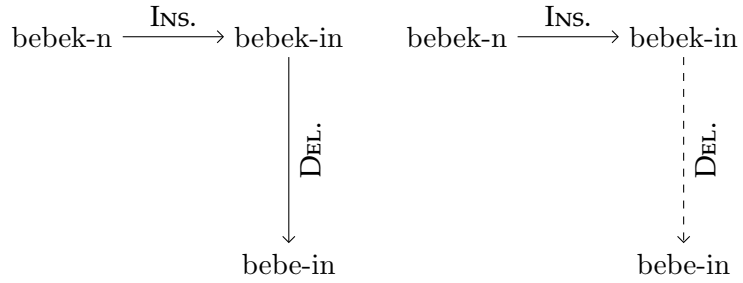
(13) Example of self-destructive feeding, data from (Kenstowicz & Kisseberth 1979)

UNDERLYING FORM	/ip-n/	/bebek-i/	/bebek-n/
∅ → i / C_C#	[ipin]	–	[bebekin]
k → ∅ / V_V	–	[bebei]	[bebein]
OBSERVED FORM	[ipin]	[bebei]	[bebein]

This phenomenon is a case of overapplication: the application of the deletion process obscures the environment in which the first process applied. What is the space of compatible

lexicons and grammars associated with this language? I present the space of joint lexicon-grammars for the seeding and counter-seeding ORs in (14), utilizing the same notation as adopted in Chapters 1 and 3.

(14) Compatible joint lexicons and grammars for the interacting seeding (LEFT) and counter-seeding (RIGHT) ORs



Contrast these hypotheses with those given for the feeding and bleeding languages derived for the revised Baković languages. The distribution observed in the seeding and counter-seeding cases closely follows that observed in the counter-bleeding language, albeit flipped vertically. The insertion input-provides deletion, as the application of insertion creates a new input that deletion can apply to that it could not before. However, unlike transparent feeding interactions, the second process, deletion, output-*removes* the first process: the application of deletion removes a possible output that insertion could have generated. Despite the interaction being formally a feeding interaction, the prediction is that the difference in the number of compatible grammars and therefore performance for the seeding versus counter-seeding interaction will be smaller than the difference in a transparent feeding and counter-feeding interaction. The model thus makes testable predictions on different process interactions beyond the traditional four-way distinction assumed thus far, allowing us to evaluate whether the effects observed in this model on replicability indeed persists on other surface patterns as well.

5.3.2 In what manner can a process interact with another?

Lastly, beyond whether or not a process can facilitate or block the application of another process is the manner in which a process interacts with the other. There are two categories along this dimension: interactions ① ON ENVIRONMENT, and ② ON FOCUS(McCarthy 1999).

Interactions on environment result from the application of one process creating an environment for which another can apply. For example, the vowel deletion process used throughout this dissertation creates the environment in which the subsequent palatalization process can apply. In contrast, interactions on focus instead result from a process

generating or eliminating the target segment of another process. For example, Low German exhibits independent processes of spirantization and devoicing (Kiparsky 1968; Kenstowicz & Kisseberth 1971). They are defined in (15).

(15) a. Spirantization: [−son, +voice] → [+cont] / V_
b. Devoicing: [−son] → −voice / _#

The application of devoicing would bleed the application of spirantization, not as a result of the environment changing, but as a consequence of changing the target to a non-applicable segment. The observed counter-bleeding on focus interaction, as well as a hypothetical bleeding on focus interaction, for these processes are given in (16) and (17).

(16) Counter-bleeding on focus interaction in Low German, taken from (Zuraw 2020)

Underlying Form	/taːg/	/taːg-ə/	/haʊz/
[−son, +voice] → [+cont] / V_	[taːɣ]	[taːɣ-ə]	–
[−son] → −voice / _#	[taːx]	–	[haʊs]
Observed Form	[taːx]	[taːɣ-ə]	[haʊs]

(17) Hypothetical bleeding on focus interaction

Underlying Form	/taːg/	/taːg-ə/	/haʊz/
[−son] → −voice / _#	[taːk]	–	[haʊs]
[−son, +voice] → [+cont] / V_	–	[taːɣ-ə]	–
Observed Form	[taːk]	[taːɣ-ə]	[haʊs]

Notably, the distributions of allophones are distinct between each interaction. In the counter-bleeding interaction, we observe the [x]~[ɣ] alternation, whereas in the bleeding interaction, we observe the [k]~[ɣ] alternation. As discussed in Chapter 2, the model denotes the similarity of two forms based on their weighted Levenshtein distance. [x] and [ɣ] differ only along a single dimension, that of voicing. In contrast, [k] and [ɣ] differ along two dimensions, that of voicing as well as continuancy. Given this toy example, then, the noisy channel would bias the model differently depending on which language is provided. Examining the effect phonological similarity has on the resultant lexicophonological space would be an interesting avenue of future research.

Bibliography

Anttila, Arto, et al. 2007. "Variation and optionality." *The Cambridge Handbook of Phonology*, 519–536.

Baković, Eric. 2011. "Opacity and ordering." *The Handbook of Phonological Theory*, 40–67.

Baković, Eric, & Lev Blumenfeld. 2018. "Overapplication conversion." *Hana-bana: A Festschrift for Junko Ito and Armin Mester.*

———. 2022. "A formal typology of process interactions." *Manuscript, University of California, San Diego.*

Bane, Max, & Jason Riggle. 2008. "Three correlates of the typological frequency of quantity-insensitive stress systems." In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, 29–38.

Benua, Laura. 1997. "Transderivational Identity: Phonological Relations between Words." PhD diss., University of Massachusetts, Amherst.

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

Boersma, Paul, et al. 1997. "How we learn variation, optionality, and probability." In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam*, 21:43–58.

Boersma, Paul, & Bruce Hayes. 2001. "Empirical tests of the gradual learning algorithm." *Linguistic Inquiry* 32 (1): 45–86.

Brooks, K Michael, Bozena Pajak, & Eric Baković. 2013. "Learning biases for phonological interactions." In *Poster presented at Annual Meetings on Phonology.*

Bybee, Joan. 2003. *Phonology and Language Use.* Vol. 94. Cambridge University Press.

Calamaro, Shira, & Gaja Jarosz. 2015. "Learning general phonological rules from distributional information: A computational model." *Cognitive Science* 39 (3): 647–666.

Chomsky, Noam. 1965. "Aspects of the theory of syntax." *Multilingual Matters: MIT Press*, 1–15.

Chomsky, Noam, & Morris Halle. 1965. "Some controversial questions in phonological theory." *Journal of linguistics* 1 (2): 97–138.

Chomsky, Noam, & Morris Halle. 1968. *The Sound Pattern of English.* MIT Press.

Ellis, Kevin, Adam Albright, Armando Solar-Lezama, Joshua B Tenenbaum, & Timothy J O'Donnell. 2022. "Synthesizing theories of human language with Bayesian program induction." *Nature Communications* 13 (1): 5024.

Ettlinger, Marc. 2008. "Input-driven Opacity." PhD diss., University of California, Berkeley.

Ettlinger, Marc, Ann R Bradlow, & Patrick CM Wong. 2014. "Variability in the learning of complex morphophonology." *Applied Psycholinguistics* 35 (4): 807–831.

Feldman, Naomi H, & Thomas L Griffiths. 2007. "A rational account of the perceptual magnet effect." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29. 29.

Frisch, Stefan A, Janet B Pierrehumbert, & Michael B Broe. 2004. "Similarity avoidance and the OCP." *Natural language & linguistic theory* 22 (1): 179–228.

Geman, Stuart, & Donald Geman. 1984. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 6, 721–741.

Gerken, LouAnn, & Alex Bollt. 2008. "Three exemplars allow at least some linguistic generalizations: Implications for generalization mechanisms and constraints." *Language Learning and Development* 4 (3): 228–248.

Gildea, Daniel, & Dan Jurafsky. 1996. "Learning bias and phonological-rule induction." *Computational Linguistics* 22 (4): 497–530.

Goldwater, Sharon, & Mark Johnson. 2004. "Priors in Bayesian learning of phonological rules." In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology: Current Themes in Computational Phonology and Morphology*, 35–42.

Goodman, Noah D, Joshua B Tenenbaum, Jacob Feldman, & Thomas L Griffiths. 2008. "A rational analysis of rule-based concept learning." *Cognitive science* 32 (1): 108–154.

Halle, Morris. 1962. "Phonology in generative grammar." *Word* 18 (1-3): 54–72.

Hansson, Gunnar Ó, & Ronald Sprouse. 1999. "Factors of change: Yowlumne vowel harmony then and now." *Proceedings of WSCLA IV*, 39–57.

Hastings, W Keith. 1970. "Monte Carlo sampling methods using Markov chains and their applications."

Heinz, Jeffrey. 2011. "Computational phonology – part I: Foundations." *Language and Linguistics Compass* 5 (4): 140–152.

Hulst, Harry van der. 2016. "Vowel harmony." In *Oxford Research Encyclopedia of Linguistics.*

Ito, Junko, & Armin Mester. 2003. "On the sources of opacity in OT: Coda processes in German." *The Syllable in Optimality Theory*, 271–303.

Jarosz, Gaja. 2006a. "Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory." PhD diss., Johns Hopkins University.

———. 2006b. "Richness of the base and probabilistic unsupervised learning in Optimality Theory." In *Proceedings of the Eighth Meeting of the ACL Special Interest Group on Computational Phonology and Morphology at HLT-NAACL 2006*, 50–59.

———. 2014. "Serial markedness reduction." In *Proceedings of the Annual Meetings on Phonology*, vol. 1. 1.

———. 2015. "Expectation driven learning of phonology." *Manuscript, University of Massachusetts Amherst*.

———. 2016. "Learning opaque and transparent interactions in Harmonic Serialism." In *Proceedings of the Annual Meetings on Phonology*, vol. 3.

Kager, René. 1999. "Surface opacity of metrical structure in Optimality Theory." *The Derivational Residue in Phonological Optimality Theory* 28:207–245.

Kaye, Jonathan Derek. 1974. "Opacity and recoverability in phonology." *Canadian Journal of Linguistics/Revue canadienne de linguistique* 19 (2): 134–149.

Kenstowicz, Michael. 1995. "Base-identity and uniform exponence: alternatives to cyclicity." *Current Trends in Phonology: Models and Methods*, 365–394.

Kenstowicz, Michael, & Charles Kisseberth. 1971. "Unmarked bleeding orders."

———. 1979. *Generative Phonology: Description and Theory*. Publisher: Academic Press.

Kim, Yun Jung. 2012. "Do learners prefer transparent rule ordering? An artificial language learning study." In *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 48:375–386. 1. Chicago Linguistic Society.

King, Robert D. 1969. *Historical Linguistics and Generative Grammar*. Prentice-Hall.

———. 1973a. "In defense of extrinsic ordering."

———. 1973b. "Rule insertion." *Language*, 551–578.

Kiparsky, Paul. 1965. "Phonological Change." PhD diss., Massachusetts Institute of Technology.

———. 1968. "Linguistic Universals and Linguistic Change." In *Universals in Linguistic Theory*, 170–202.

———. 1971. "Historical Linguistics." In *A Survey of Linguistic Science*, 576–642.

———. 2000. "Opacity and cyclicity." *The Linguistic Review* 17:351–365.

Kirov, Christo. 2017. "Recurrent neural networks as a strong domain-general baseline for morpho-phonological learning." In *Poster presented at the 2017 Meeting of the Linguistic Society of America*.

Kirov, Christo, & Ryan Cotterell. 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.

Kisseberth, Charles W. 1973. "The interaction of phonological rules and the polarity of language."

Levy, Roger. 2008. "A noisy-channel model of human sentence comprehension under uncertain input." In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, 234–243.

Magri, Giorgio. 2012. "Convergence of error-driven ranking algorithms." *Phonology* 29 (2): 213–269.

Martin, Alexander, & James White. 2021. "Vowel harmony and disharmony are not equivalent in learning." *Linguistic Inquiry* 52 (1): 227–239.

McCarthy, John. 1999. "Sympathy and phonological opacity." *Phonology* (January): 331–399.

———. 2000. "Harmonic serialism and parallelism." In *North East Linguistics Society*, 30:8. 2.

———. 2007. *Hidden Generalizations: Phonological Opacity in Optimality Theory.* Equinox.

Moreton, Elliott. 2008. "Analytic bias and phonological typology." *Phonology* 25 (1): 83–127.

Moreton, Elliott, & Joe Pater. 2012. "Structure and substance in artificial-phonology learning, part II: Substance." *Language and Linguistics Compass* 6 (11): 702–718.

Nazarov, Aleksei, & Joe Pater. 2017. "Learning opacity in stratal maximum entropy grammar." *Phonology* 34 (2): 299–324.

Nelson, Max. 2019. "Segmentation and UR acquisition with UR constraints." *Proceedings of the Society for Computation in Linguistics* 2 (1): 60–68.

O'Bryan, Margie. 1974. "Opacity and Rule Loss."

Ohala, James J. 1992. "What's cognitive, what's not, in sound change." In *Diachrony within Synchrony: Language History and Cognition. Duisberger Arbeiten zur Sprach-und Kulturwissenschaft 14*, 309–355. Frankfurt: Peter Lang.

Paster, Mary. 2005. "Subcategorization vs. output optimization in syllable-counting allomorphy." In *Proceedings of the 24th West Coast Conference on Formal Linguistics*, 24:326. Cascadilla Proceedings Project Somerville.

Pater, Joe, Robert Staubs, Karen Jesney, & Brian Cantwell Smith. 2012. "Learning probabilities over underlying representations." In *Proceedings of the 12th Meeting of the Special Interest Group on Computational Morphology and Phonology*, 62–71.

Peperkamp, Sharon, Rozenn Le Calvez, Jean-Pierre Nadal, & Emmanuel Dupoux. 2006. "The acquisition of allophonic rules: Statistical learning with linguistic constraints." *Cognition* 101 (3): B31–B41.

Perkins, Laurel, Naomi Feldman, & Jeffrey Lidz. 2017. "Learning an input filter for argument structure acquisition." In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2017)*, 11–19.

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

Prickett, Brandon, & Gaja Jarosz. 2021. "Modeling the Acquisition of Phonological Interactions: Biases and Generalization." In *Proceedings of the Annual Meetings on Phonology*, vol. 8.

Rafferty, Anna N, Thomas L Griffiths, & Marc Ettlinger. 2013. "Greater learnability is not sufficient to produce cultural universals." *Cognition* 129 (1): 70–87.

Rasin, Ezer, Iddo Berger, Nur Lan, & Roni Katzir. 2015. "Learning rule-based morphophonology." *Manuscript, Tel Aviv University.*

Riggle, Jason. 2010. "Sampling rankings." *ROA-1075.*

Sanders, Robert Nathaniel. 2003. "Opacity and Sound Change in the Polish Lexicon." PhD diss., University of California, Santa Cruz.

Schneider, Jordan, Laurel Perkins, & Naomi H Feldman. 2020. "A noisy channel model for systematizing unpredictable input variation." In *Proceedings of the 44th Annual Boston University Conference on Language Development*, 533–547.

Smolensky, Paul. 1995. "On the internal structure of the constraint component Con of UG." *ROA-86.*

———. 1996. "The initial state and 'richness of the base' in Optimality Theory." *ROA-293* 293.

Smolensky, Paul, & A Prince. 1993. *Optimality Theory: Constraint interaction in generative grammar.* Publisher: Wiley Online Library.

Stanton, Juliet. 2016. "Learnability shapes typology: the case of the midpoint pathology." *Language* 92 (4): 753–791.

Staubs, Robert D. 2014. "Computational modeling of learning biases in stress typology." PhD diss., University of Massachusetts, Amherst.

Steriade, Donca. 2000. "Paradigm uniformity and the phonetics-phonology boundary." *Papers in Laboratory Phonology* 5:313–334.

Sutskever, Ilya, Oriol Vinyals, & Quoc V Le. 2014. "Sequence to sequence learning with neural networks." *Advances in Neural Information Processing Systems* 27.

Wagner, Robert A, & Michael J Fischer. 1974. "The string-to-string correction problem." *Journal of the ACM (JACM)* 21 (1): 168–173.

White, James. 2017. "Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias." *Language*, 1–36.

Wilson, Colin. 2006. "Learning phonology with substantive bias: An experimental and computational study of velar palatalization." *Cognitive science* 30 (5): 945–982.

Yang, Charles. 2005. "On productivity." *Linguistic Variation Yearbook* 5 (1): 265–302.

Yang, Christopher, & Kevin Ellis. 2021. "Phonological Interactions, Process Types, and Minimum Description Length Principles." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43. 43.

Zuraw, Kie. 2020. "Process Interactions II." *Handout, University of California, Los Angeles.*