Title
Phonological markedness effects in reanalysis

Permalink
https://escholarship.org/uc/item/5pg8z3x9

Author
Kuo, Jennifer

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Phonological markedness effects in reanalysis

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Linguistics

by

Jennifer Kuo

2023

ABSTRACT OF THE DISSERTATION

Phonological markedness effects in reanalysis

by

Jennifer Kuo

Doctor of Philosophy in Linguistics

University of California, Los Angeles, 2023

Professor Bruce P. Hayes, Chair

Paradigms with conflicting data patterns can be difficult to learn, resulting in acquisition error. In this dissertation, I look at how paradigms are reanalyzed over time to gain insight into the factors that influence morphophonological learning. Existing models of morphophonology (e.g. Hare & Elman 1995; Albright 2002b,a, 2010) predict reanalysis to be frequency-matching, occurring in a way that matches probabilistic distributions within the paradigm. I propose that in fact, reanalysis responds to two factors: both frequency-matching and a bias towards less marked outputs. Additionally, markedness effects in reanalysis are argued to be restricted to so-called 'active' markedness effects, which are already present in the language as stem phonotactics.

I present three case studies, all from Austronesian languages, where reanalysis is arguably sensitive to a markedness bias, and confirm this by implementing a quantitative model of reanalysis. This model, outlined in Chapter 2, simulates the cumulative effect of reanalyses over time with an iterated learning paradigm. In each iteration, learning is modeled using Maximum Entropy Harmonic Grammar (MaxEnt; Smolensky 1986; Goldwater & Johnson 2003), with a markedness bias implemented as a Gaussian prior (Wilson 2006).

The three studies are presented in Chapters 3-5. All three cases involve paradigms where there is ambiguity in how the suffixed forms will surface, resulting in reanalysis of these suffixed forms. The first case study concerns Malagasy weak stems; frequency-matching models predict reanalysis towards one alternant, but instead there has been reanalysis towards another statistically dispreferred alternant. I argue that this outcome is motivated by avoidance of intervocalic stops, and show that this analysis does better than alternative explanations.

The second and third case studies concern Samoan and Māori. In both languages, certain suffixes have multiple allomorphs with an unpredictable distribution. In Samoan (Chapter 4), reanalysis is generally towards the suffix allomorph predicted by frequency-matching models, but is also modulated by OCP-place effects (McCarthy 1988, 1994). Specifically, suffixed forms which violate OCP-place are more likely to be reanalyzed. In Māori, reanalysis is towards a suffix allomorph that is not predicted by frequency-matching models. I argue that reanalysis has instead been motivated by avoidance of both vowel hiatus and heavy syllables .

All three languages show evidence of reanalysis that is sensitive to a markedness bias. Moreover, all three cases are also consistent with the principle of active markedness, as the markedness effects found in reanalysis are already present in the language-specific phonotactics. Based on these results, I argue for a richer model of reanalysis in which phonotactic principles serve as a learning bias.

The dissertation of Jennifer Kuo is approved.

David Michael Goldstein

Claire Moore-Cantwell

Kie Ross Zuraw

Bruce P. Hayes, Committee Chair

University of California, Los Angeles

2023

To my family

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I had a wonderful and stimulating time at UCLA, and I owe a great deal of that to my advisors, colleagues, and friends.

First and foremost, I want to thank my advisor Bruce Hayes, who has guided me since my first year and really served as a model of the kind of linguist, teacher, and mentor that I want to be. Bruce is an incredible advisor who was able to help me see the big picture when I was stuck in the weeds of my own research, and always ready with some insightful comments. He is also endlessly encouraging and enthusiastic, and this dissertation could not have happened without his guidance and support.

I am also very thankful Kie Zuraw, who was generous with her time, providing me with some much-needed sensible advice and moral support. She was the one who taught me to treat research writing as creative writing. Thanks also to Claire Moore-Cantwell for helping me to think about my work from a fresh perspective, and for reminding me to take a break during the long and seemingly endless process of dissertation writing. Thanks to David Goldstein, who graciously agreed to join my committee and brought valuable insight from the perspective of historical linguistics. Thanks also to Megha Sundara, Sun-Ah Jun, Pat Keating, Ben Eischens, Shu-hao Shih, Matt Faytak, and other faculty for your teaching, fun conversations, and helpful advice.

Thanks to the linguistics community at UCLA, who always made me feel welcome. First, thanks to my graduate cohort at UCLA: Jinyoung, Jake, Colin, Noah, Mia, Phill, Angelica, Joy, and Tyler. I don't know how I would've made it through that intimidating first year without you all, and I hope we keep in contact as we go our ways. Thanks in particular to Jinyoung, my first friend in LA, who is in many ways wiser (and more organized) than I am. Thanks also to Canaan, Z.L., Hiro, Meng, Christine, Jahnavi, Connor, Jeremy, Jian-Leat, and many other graduate students for your camaraderie and fun conversations. I also benefited greatly from the members of the Phonology Seminar, both from the stimulating research that was shared there, and from the various pieces of advice

2020	MA in Linguistics
	University of California, Los Angeles
2018	BA in Linguistics (minor in Japanese)
	Dartmouth College

PUBLICATIONS

Elkins, Noah and Kuo, Jennifer. (2023). A prominence account of the Northern Mam weight hierarchy. *Supplemental Proceedings of the 2022 Annual Meeting on Phonology*. https://doi.org/10.3765/amp.v10i0.5433

Grabowski, Emily and Kuo, Jennifer. (2023). Comparing K-means and OPTICS clustering algorithms for identifying vowel categories. *Proceedings of the Linguistic Society of America, 8*(1), 5488. https://doi.org/10.3765/plsa.v8i1.5488

Kuo, Jennifer. (2023). Evidence for prosodic correspondence in the vowel alternations of Tgdaya Seediq. *Phonological Data and Analysis, 5*(3), 1-31. https://doi.org/10.3765/pda.v5art3.77

Kuo, Jennifer. (2018). A large-scale smartphone-based sociophonetic study of Taiwan Mandarin. *Asia-Pacific Language Variation, 4*(2), 197–230.

CHAPTER 1

Introduction

Recent developments in phonology show that people can learn fine-grained, probabilistic generalizations about their language (e.g. Bybee & Moder 1983; Prasada & Pinker 1993; Albright & Hayes 2003). A central debate in research on phonological learning concerns how this type of statistical learning interacts with other language-specific learning biases. One view holds that phonological learning can be explained by a domain-general ability for statistical learning. In fact, statistical learning has a lot of explanatory power and is often sufficient to explain speakers' phonological intuitions (Section 1.1.1). On the other hand, there is also growing evidence that learning can be constrained by various synchronic learning biases (analytic biases; Moreton 2008). Such biases are argued to be specific to the language faculty, or at least grounded in human cognitive mechanisms, and therefore reflect principles of Universal Grammar (UG).

In recent years, extensive experimental work has been done to tease apart the effects of statistical learning and competing biases, including studies of child language acquisition and adult nonce-word experiments (i.e. wug tests). Since Kiparsky (1965, 1997, 1978, et seq.), it has also been recognized that *language change* serves as a robust "natural laboratory" for understanding how children learn and mislearn patterns outside the constraints of a laboratory setting.

In this dissertation, I adopt the latter approach and probe into how statistical learning interacts with learning biases by looking at a type of language change called **reanalysis**, where morphophonological paradigms are remade over time. Existing models of reanalysis (and more generally of morphophonological learning) predict that learners rely only

on statistical regularities within a paradigm. As a preview of the results, I find that reanalysis in three languages—Malagasy, Samoan, and Māori—cannot be explained entirely by this type of local statistical learning, and is also sensitive to markedness effects *external* to the paradigm. I also argue for a restricted view of markedness bias, where markedness effects present in reanalysis must already be active in the stem phonotactics of a language.

1.1 Frequency-matching vs. learning biases

1.1.1 Frequency-matching

In phonological learning, speakers are known to use statistical properties of the input data to make predictive generalizations. In particular, when speakers are faced with variable patterns in a morphophonological paradigm, they **frequency-match**, applying these patterns in a way that matches the proportion at which they occur in the data. For example, Ernestus & Baayen (2003) study Dutch final devoicing, where final obstruents are devoiced, results in voicing alternations such as in (1).

(1) *Dutch voicing alternations*
     [vɛrˈʋɛit]   [vɛrˈʋɛi**d**en]   'widen'   (t∼d)
     [vɛrˈʋɛit]   [vɛrˈʋɛi**t**en]   'reproach'   (non-alternating)

When we look at the Dutch lexicon, rates of voicing alternation are partially predictable from statistical tendencies. For example, final [p] is non-alternating around 90% of the time in the lexicon. Conversely, final [f] alternates with [v] around 70% of the time. Ernestus & Baayen (2003) find that when speakers are told to provide the suffixed form of wug stems, they apply voicing alternations in a way that aggregately matches these distributional patterns.

Frequency-matching has been found to predict adult linguistic behavior in various other experiments, including: Eddington (1996, 1998, 2004); Coleman & Pierrehum-

bert (1997); Berkley (2000b); Zuraw (2000); Bailey & Hahn (2001); Frisch & Zawaydeh (2001); Albright (2002b); Albright & Hayes (2003); Hayes & Londe (2006); Hayes et al. (2009); Pierrehumbert (2006); Jun & Lee (2007). Sociolinguistic studies also demonstrate that children frequency-match adult speech patterns (Labov 1994, Ch. 20).

Note that while there is extensive evidence for frequency-matching in adults, there is also evidence that children *over-regularize* patterns instead of frequency-matching (Hudson Kam & Newport 2005, 2009; Schumacher & Pierrehumbert 2021). Experimental evidence shows that when learners overgeneralize, they tend to do so towards the more frequent pattern, but sometimes also converge towards the minority pattern (Schumacher & Pierrehumbert 2021).

Why might both frequency-matching and overregularization be observed? One possible explanation is that the choice between the two depends on input size (or, the size of the learner's lexicon). That is, when given very little evidence for a pattern, learners initially overregularize gradient patterns. As they receive more input, they become better at frequency matching. Various studies have found empirical support for this learning trajectory (e.g. Levelt et al. 2000; Gnanadesikan 2004; Jarosz 2010).

Notably, frequency-matching predicts change to be preservatory (i.e. maintaining the statistical distributions of the input), while overregularization predicts reanalysis towards the more frequent variant. However, where reanalysis is not predicted by statistical distributions within a paradigm, neither approach can provide a complete picture of the factors driving reanalysis.

1.1.2 Learning biases

In many cases, such as with the Dutch voicing alternations discussed above, frequency-matching (and more generally statistical learning) is sufficient to explain speakers' phonological intuitions. On the other hand, there is growing evidence that learning is also constrained by various learning biases. Evidence for bias comes from cases where speakers

fail to frequency-match, and instead over-learn patterns (reflecting a bias for the target pattern; e.g. Kuo 2023) or under-learn them (reflecting a bias against the target pattern; e.g. Hayes et al. 2009). Broadly speaking, two types of bias are seen. These are i) synchronic preferences towards acquiring some phonological patterns over others (analytic bias; Moreton 2008), and ii) channel bias, or factors in the production or perception of speech that affect the faithful transmission of speech sounds over time.

One view of biased phonological learning, which Hayes et al. (2009) call the 'strong UG' approach, holds that learners have an innate universal constraint set (or a set of rules). Phonological patterns that cannot be derived by these constraints are simply unlearnable. This approach is taken up by Becker et al. (2011), who use a nonce-word study to test whether Turkish speakers have learned the lexical statistics of consonant laryngeal alternations (examples in (2)).

Corpus studies show that word length, place of articulation, and preceding vowel quality (height/backness) are all significant predictors of rates of laryngeal alternation in Turkish. In a wug test, Turkish speakers were found to be sensitive to length and place of articulation, but not preceding vowel quality.

Becker et al. (2011) argue that this is because constraints which encode interactions of vowel height/backness with consonant voicing are not available to learners (i.e. not in the universal set of constraints), making the Turkish vowel-consonant interactions unlearnable.

(2)　*Turkish laryngeal alternations (Becker et al. 2011, p. 85)*

	BARE STEM	POSSESSIVE	
Non-alternating	at**tʃʰ**	at**tʃʰ**-ɨ	'hunger'
	ana**tʃʰ**	ana**tʃʰ**-ɨ	'female cub'
tʃ~dʒ	gy**tʃʰ**	gy**dʒ**-ɨ	'force'
	ama**tʃʰ**	ama**dʒ**-ɨ	'target'

In contrast to this stronger approach, I assume that learning biases are a 'soft' preference: linguistically unnatural patterns are harder to learn, but can still be learned given

sufficient evidence. I adopt the soft bias approach for several reasons: typologically, there is substantial evidence that phonological patterns are not always natural; some examples can be found in Hellberg (1978),Bolognesi (1998), Ito & Mester (2003), Hansson (2007), and Odden (2007). Experimental work on bias learning has also shown that speakers can learn unnatural patterns. For example, Hayes et al. (2009) conduct a wug test on Hungarian vowel harmony and find that when given enough evidence, speakers can learn unnatural patterns present in the lexicon. Artificial Grammar Learning experiments also suggest that both infants and adults can learn unnatural phonological patterns; see for example Seidl & Buckley (2005) on infant learning and Wilson (2006) and White (2013, 2017) on adult learning.

Two types of bias have been discussed in the literature: 1) complexity bias, or a bias against formally complex patterns (Moreton & Pater 2012a) and 2) substantive bias, or a bias against phonetically unnatural patterns, where phonetic naturalness includes factors such as perceptual similarity and articulatory ease (Moreton & Pater 2012b).

Within the literature on substantive bias, most recent work has focused on perceptual similarity bias, or a preference for alternation patterns that involve perceptually smaller changes (e.g. Steriade 2001; Wilson 2006; White 2013; Glewwe 2019). For example, in an AGL study, Wilson (2006) found that participants trained to palatalize velars before [e] generalized palatalization to apply before [i] (/ke/∼[tʃe] generalizes to /ki/∼[tʃi]). On the other hand, those trained to palatalize velars before [i] did not generalize to the [e] context (/ki/∼[tʃi] ↛ /ke/∼[tʃe]). This asymmetry is argued to be rooted in a perceptual similarity bias: speakers prefer [k]∼[tʃ] alternation before [i] because [k] and [tʃ] are perceptually more similar before [i] than before [e].

Another possibility, which is the empirical focus of this dissertation, is a so-called **markedness bias**, or a bias against output forms that are harder to produce or perceive. Whereas perceptual similarity bias looks at the mapping between related forms (e.g. input-output, output-output), markedness bias targets the surface output form.

For example, intervocalic stops are cross-linguistically dispreferred and often the tar-

get of lenition processes at morpheme boundaries (e.g. Jun 1994 on Korean; Hayes 2011 on English; Wheeler 2005 on Catalan). They are also argued to be dispreferred from an articulatory point of view (Kirchner 1998). It is therefore conceivable that learners are biased against output forms like [pati] relative to [paɾi], where the former is dispreferred because it has an intervocalic [t].

1.2   Reanalysis and modeling reanalysis

One way to understand the interaction between frequency effects and bias is to see how learners deal with conflicting data patterns in morphophonological paradigms. For example, Albright & Hayes (2003) look at past tense formation in English, where there are sometimes conflicting generalizations on how to derive the past tense for a given word.

The data in (3) gives an illustrative example; the generalizations given here come from Albright & Hayes (2003), who use data from the CELEX database (Baayen et al. 1996). In this case, a hypothetical nonce word *gleed* [glid] has at least four possible past tense forms, each associated with statistical generalizations of varying strength. In other words, to form the past tense for *gleed*, learners must choose one of multiple competing options.

(3)   Rules for forming the past tense of *gleed* (Albright & Hayes 2003, p. 128)[1]

Output	Rule	Adjusted Confidence (p̂)	Examples
gleeded	$\emptyset \rightarrow$ əd / {d, t}__	0.872	*want, need, start, decide*
gled	i$\rightarrow$ ɛ / {l,r}__d	0.793	*read, lead, bleed, breed*
glode	i$\rightarrow$ o / C__[+cons]	0.033	*speak, freeze, weave*
gleed	no change / __{d,t}	0.024	*shed,spread,put*

More generally, given a bare stem, if a learner has never heard the past tense form, they must choose between several possible options. Albright & Hayes (2003) (and many other studies) probe at how learners resolve this type of paradigm ambiguity using nonce-word experiments. Another way to understand speaker intuitions is to see how they mislearn *real words*. In particular, the challenge posed by conflicting data patterns can cause children to mislearn words. For example, children often mislearn the past tense of *go* as *goed* (cf. *went*).

Reanalysis, which is my empirical focus, refers to when innovative variants such as *goed* are adopted into speech communities and passed down to the next generation of speakers. In essence, ambiguity in one slot of a paradigm causes other slots of the paradigm to be reanalyzed, resulting in a type of language change. Some examples of reanalysis in English past tense are given in (4).[2]

(4)   *Reanalyses in English past tense formation*

STEM   PST
help   halp→helped (1300, OED)
dive   dived→dove (1800, OED)

_____

[1]Generalizations here are based off of the Albright & Hayes (2003) Minimal Generalization Learner (MGL), though in principle similar generalizations could be found using other approaches. Note that the MGL calculates an adjusted confidence value, which accounts for not just how accurate a rule is, but also how much evidence there is for the rule (i.e. the number of forms the rule could apply to). For simplicity, this table presents only adjusted confidence.

[2]In some cases, such as *dived* vs. *dove*, there is still variation, and both variants are observed.

Reanalysis is closely related to *analogical change.* Analogy describes when a word is changed on the basis of perceived similarity to another word, but has also been more broadly defined as any changes to a word that cannot be explained by regular sound changes. In fact, much work on change in morphophonological paradigms uses the term analogy (e.g. Hare & Elman 1995; Albright & Hayes 2003; Albright 2008, and many more). I use the term reanalysis to emphasize that I am focusing on systematic changes shared by multiple words in the same paradigmatic relationship. In contrast, analogy has traditionally been used to describe associations between individual words.

Moreover, in probabilistic models of morphophonology, the term 'analogy' also has theoretical implications, and can refer to specific exemplar-based implementations (e.g. Analogical Modeling of Language; Skousen 1989) which contrast with rule-based (or constraint-based) frameworks. The term reanalysis is more neutral, and less tied to exemplar-based frameworks.

The rest of this section discusses early work on analogical change and ties this to more recent work that focuses on modeling reanalysis in probabilistic grammars.

1.2.1 Analogy as associative proportions

Historical linguists of the late 19th century defined analogy as an associative process in language change, where individual words become the model for change in other words. For example, a famous analogy in the history of Latin eliminated a stem-final contrast between [r] and [s] (Hock, p. 179-190; Kiparsky 1997; Albright 2005, etc), as shown in (5). This process of analogical leveling was described in terms of the four-part proportional analogy in (6), where words like [honoːs] were influenced by non-alternating stems like [soror] 'sister' (Hock 1991, pp. 179-190).

(5)   *Pre-leveling* *Post-leveling*
      hono:**s**   hono**r**
      hono:**ris**   hono:**ris**
      hono:**ri:**   hono:**ri:**
      hono:**rem**   hono:**rem**

(6)   [soro:ris]:[soror] :: [hono:ris]:[honor]

Analogy was used to account for apparent irregularities in language change (in contrast to regular sound change). As such, although it was generally recognized that analogical changes results from the collective influence of many words, analogy was often discussed in terms of individual cases. Scholars of the time pointed out various tendencies about the direction of analogy. For example, Kuryłowicz & Winters (1947) and Mańczak (1957) listed guiding principles on the direction of analogical change, citing factors like markedness and frequency. However, there wasn't much work on analogical change as a probabilistic phenomenon, and guidelines did not make concrete, language-specific predictions about the direction and outcome of analogy.

1.2.2 Analogy (and reanalysis) as regularization

Kiparsky's (1965; 1988; 2012, et seq.) seminal works redefined analogy as regularization, or the reduction of unmotivated grammatical complexity and idiosyncrasy. More concretely, the direction of analogical change is governed by principles such as bleeding/feeding ordering and reduction of rule opacity. Under this approach, analogy is closer to what I refer to as reanalysis, encompassing not just exemplar-based changes (i.e. associative proportions), but also systematic changes which require consideration of the grammar and lexicon as a whole.

Kiparsky's approach also shifted the focus away from looking at analogy as a purely diachronic phenomenon. Instead, diachronic changes and synchronic morphophonology

are enforced using the same principles. In subsequent years, there have been various attempts to formalize analogy/reanalysis in a more systematic way. The idea of reanalysis as regularization (or more specifically the idea that paradigms tend to be uniform) has been formalized in Optimality Theory as constraints like PARADIGM UNIFORMITY, LEVEL, and UNIFORM EXPONENCE (Kenstowicz 1996, 1997; Steriade 1997; Kager 2000, and more). For example, Kenstowicz (1996) analyzes the [hono:s] > [honor] change as the promotion of a constraint which requires uniformity in noun paradigms.

Kiparsky formalizes the idea of regularization in language change using constraint-based frameworks, but he notes that the same arguments can be made in rule-based frameworks. In fact, subsequent work, such as by Dresher (1980; 1985; 2015, etc.), argues for an approach in which both diachronic change and synchronic grammars are viewed through the framework of rule-based generative phonology.

For example, Lahiri & Dresher (1999) use this rule-based approach to explain seeming inconsistencies in the history of English vowel length alternations. Middle English underwent two processes, Open Syllable Lengthening (OSL) and Trisyllabic Shortening (TSS), defined in (7) and (8), which interacted to result in vowel length alternations. However, these processes are not reflected consistently in present day English; there is considerable variation the length of vowels that should have undergone OSL and TSS (Minkova 1982). Moreover, OSL is expected to have resulted in vowel length alternations in English singular/plural nominal paradigms, but such alternations are largely absent from present-day English.

(7)   Open Syllable Lengthening (OSL): A short stressed vowel in an open syllable must be long.

(8)   Trisyllabic Shortening (TSS): A long stressed vowel followed by two unstressed syllables must be short.

Lahiri & Dresher (1999, p. 698) propose that this is because OSL interacted with another process–the loss of schwas in inflected forms–resulting in an opaque alternation pattern. This is demonstrated in §1.2.2; what used to be a phonologically regular alternation became unpredictable. The addition of plural /s/ variably leaves the vowel length unchanged, shortens a long vowel, and even lengthens a short vowel. Learners, faced with this opacity, are argued to have leveled the length alternation.

(9) Expected singular-plural pairs in Middle English, from Dresher (2015, p. 28)

Before loss of schwa			After loss of schwa	
singular	plural		Singular	Plural
stōn	stōnes	(OSL)	stōn	stōns
bōdi	bodies	(TSS)	bōdi	bodis
god	gōdes	(OSL)	god	gōds
bēver	beveres	(TSS)	bēver	bevers

In this example, a better understanding of English length alternations was achieved by considering how lengthening and shortening interacted with other phonological processes in the synchronic grammar. In general, like Kiparsky's approach to analogical change, this generative analysis put a greater focus on reanalyses not as isolated processes, but as motivated by grammar-wide principles. Moreover, the division between regular sound change and analogy is less drastic, as both are formulated in terms of generative rules.

1.2.3   Probabilistic models of reanalysis

Recent research on reanalysis/analogy has become increasingly centered around capturing speakers' detailed statistical knowledge about regularities (and subregularties) in morphophonology. Various computational implementations of reanalysis have been developed to explain this type of statistical knowledge.

Formal implemented models of reanalysis face two challenges. First, they must be powerful enough to capture gradient and probabilistic data. In particular, experimental

evidence suggests that people can make very fine-grained generalizations about morphophonological alternations (Bybee & Moder 1983; Prasada & Pinker 1993; Albright & Hayes 2003). Going back to English past tense formation as an example, sub-generalizations such as "i → ʌ / __ŋ]$_{[+\text{past}]}$" (corresponding to *fling/flung, sting/stung,* etc) co-exist with a more general suffixation rule.

On the other hand, models of reanalysis should also be restrictive. Evidence from language change and child errors show that attested reanalyses only account for a small fraction of the logically possible changes (e.g. Simões & Stoel-Gammon 1979; Clahsen et al. 2002; Kang 2006; Albright 2010). A model of reanalysis should be able to explain why speakers don't generalize some properties of the input data.

Most quantitative models of reanalysis have focused on the first challenge of capturing probabilistic patterns in the data. These models include neural networks (Rumelhart & McClelland 1987; MacWhinney & Leinbach 1991; Daugherty & Seidenberg 1994; Hare & Elman 1995), Analogical Modeling of Language (AML; Skousen 1989), symbolic analogical models (Tilburg Memory-Based Learner Daelemans et al. 2004), the Generalized Context Model (Nosofsky 1990, 2011), and decision-tree-based models (Ling & Marinov 1993). For example, Hare & Elman (1995) use a connectionist model (essentially a shallow neural network) to model reanalysis of English past tense inflection, which changed from a highly complex system in Old English to the more regular system found in today's English.

1.2.4 Probabilistic rule-based models (Albright 2002b)

While earlier work focused on developing models powerful enough to capture gradient subregularities in morphophonology, Albright (2002b,a, 2008, 2010, et seq) focuses on the second problem of developing a sufficiently restrictive model. In my work, I adopt Albright's assumptions, with some differences that will be pointed out here and expanded on in Chapter 2.

STRUCTURED SIMILARITY. This is the idea that word-forms must share the same structural description to be the basis for reanalysis. Structured similarity contrasts with exemplar-based views of reanalysis/analogy, in which any form that shares some similarity with the target word can be the basis of reanalysis (even if they don't all share the same structural description).

Albright formalizes the requirement for structured similarity by using a rule-based framework. For example, a rule like "ɪ → ʌ / __ŋ]$_{[+\text{past}]}$" describes past-tense formation in singular-past pairs like *fling/flung, sting/stung, cling/clung.* Crucially, it also captures the fact that these word-pairs all share the same structural description, where the bare stem ends in [ɪŋ].

In the rest of this dissertation, I also assume structured similarity, but encode this using constraints (in variants of Optimality Theory; Prince & Smolensky 1993). As will be further discussed in Chapter 2, I adopt an Optimality Theoretic approach because I am looking at markedness effects which are essentially generalizations about the output (product-oriented generalizations), rather than about input-output mappings. Product-oriented generalizations are easily captured using markedness constraints in OT, but are less straightforwardly accounted for in rule-based frameworks.[3]

SINGLE-BASE HYPOTHESIS. Albright also proposes that reanalysis is always from *one* paradigm slot, and it is the maximally informative one. This paradigm slot serves as the so-called 'base' of reanalysis. In cases where two slots are of similar informativeness, other factors like token frequency and a preference for morphologically simple bases might come into play (Albright 2008).

In my dissertation, I am looking at case studies where reanalysis can only occur from one paradigm slot, so in some sense, this restriction for a single base is vacuously satisfied. However, most of the cases I look at actually involve reanalysis from the *less informative*

---

[3]In reality, work by Bybee and colleagues suggests that speakers form both source-oriented generalizations (which capture input-ouput relations) and product-oriented generalizations (Bybee & Slobin 1982; Bybee & Moder 1983; Bybee 2003).

base (i.e. the one where contrasts have been neutralized). This suggests that base informativeness may be less of a hard restriction, but rather one of the many factors that affects the direction of reanalysis.

Notably, all the models discussed so far assume reanalysis to be in the direction of the statistically most probably outcome, given the distribution of sounds within a paradigm. In other words, they are all frequency-matching. However, I am arguing, on the basis of the case studies discussed below, that statistical distributions alone is not always sufficient for predicting the output of reanalysis. Instead, models of reanalysis must also account for markedness effects, or a pressure to reduce the markedness of output forms.

1.3 Active vs. universal markedness

The term "unmarked" has a broad meaning, and has been used to describe output forms that are simpler, more common, easier to produce, acquired earlier, etc. In general, markedness has come to refer to the universals of language (e.g. Jakobson 1963; Greenberg 1966), determined by Universal Grammar (Chomsky & Halle 1968; Kean 1975, and many others following them).

When we consider markedness effects in reanalysis, it is also important to consider how such effects are constrained—in other words, what is the range of markedness effects that are able to influence reanalysis (and more generally, morphophonological learning)? One view, which I refer to as "universal markedness", is that all possible markedness constraints as defined by Universal Grammar can affect reanalysis. Another view, which I call "active markedness',' is more restrictive, and predicts that markedness constraints can only affect reanalysis if they are already active in the lexicon in the form of stem phonotactics.

The active markedness proposal is attractive because it ties into existing theories of acquisition and empirical findings about the relationship between phonotactics and morphophonology. Typologically, similar phonological generalizations tend to hold within

morphemes and across morpheme boundaries; in other words, alternations are consistent with stem phonotactics (Chomsky & Halle 1968; Kenstowicz 1996). This is especially true once we consider gradient effects; Chong (2019) shows that even in cases of apparent mismatch between phonotactics and alternations, there is often some gradient phonotactic support for an alternation pattern. Additionally, alternations that are not supported by phonotactics tend to be under-attested.

Theoretically, a tight connection between phonotactics and morphophonological is built into classical Optimality Theory, where within-morpheme and cross-morpheme generalizations are modelled using the same mechanism (i.e. the same markedness constraint and constraint ranking). Another related view, held by many theories of learning and acquisition, is that phonotactics and alternations tend to be closely related because phonotactics are learned earlier, and aid in the later learning of alternations (Hayes 2004; Jarosz 2006; Tesar & Prince 2003; Yang 2016).

There is also some experimental work supporting the idea that phonotactics aids in alternation learning. For example, Pater & Tessier (2005) find that English speakers learn a novel alternation pattern better when it is supported by English stem phonotactics. In an AGL experiment, Chong (2021) trains speakers both a novel phonotactic pattern and novel alternation patterns. Results suggest that speakers draw on phonotactics to resolve ambiguities in morphophonological alternations. There is also work showing that phonotactics are easier to acquire than alternations; phonotactic generalizations are acquired earlier by children (e.g. Zamuner 2006), and can be acquired by adults even with limited input (Oh et al. 2020).

In work on compound formation, Martin (2011) also finds similar effects of active markedness. In particular, Martin presents evidence from Navajo and English that the same phonotactic constraints present within morphemes are also active in compound formation, albeit as a weaker, gradient effect. In other words, there is evidence that speakers generalize phonotactic constraints across morpheme boundaries. Given Martin's findings, it is conceivable that stem-internal phonotactics could also constrain cross-morpheme al-

ternation patterns.

Within work on language change, findings from Garrett (2008) support the idea that markedness-motivated paradigm reanalyses are a product of language-specific factors rather than a direct manifestation of UG. While Garrett's focus is on semantic (rather than phonological) markedness patterns, his findings still provide support for the idea that reanalysis is driven by markedness effects already present in the language.

For these reasons, I propose that markedness bias is restricted to active markedness effects. In other words, speakers utilize markedness principles already present in the language's phonotactics when resolving ambiguities in an alternation pattern. I will show that the case studies presented in this dissertation are all consistent with the active markedness proposal.

1.4   When can markedness-driven reanalysis occur?

My proposal, broadly speaking, is that reanalysis should be phonologically optimizing. The active markedness approach, in particular, predicts that reanalysis will result in a close correspondence between stem-internal phonotactics and cross-morpheme alternations. I also argue that this type of markedness-driven reanalysis only comes into play when there is *uncertainty* in an alternation pattern. In other words, markedness effects in reanalysis are only observed when there is conflicting evidence for which alternant should surface, and one alternant is less marked than the competing alternants.

This distinction is important because it allows mismatches between phonotactics and alternations to persist if an alternation pattern is predictable. There is crosslinguistic evidence that phonotactics-alternation mismatches can persist in a language. For example, Turkish vowel harmony operates within stems but not across compounds or phonological words (Kabak & Vogel 2001); see also Gouskova (2018) for an overview of similar mismatches. Experimental evidence from Gallagher et al. (2019) also supports the idea that speakers are able to learn different cross-morpheme and morpheme-internal phonotactic

generalizations.

Relatedly, morphophonological patterns which are not phonologically optimizing can also persist if the relevant pattern is predictable. In particular, there is crosslinguistic evidence for *phonologically conditioned suppletive allomorphy*, or cases where allomorphy has clear phonological conditioning but is not output-optimizing (Paster 2005, 2009). For example, in Tzeltal, the perfective allomorph that surfaces (*-eh* vs. *-oh*) depends on how many syllables the stem has, in a way that is not output-optimizing.

In summary, although my proposal of markedness-driven reanalysis predicts a strong connection between within-morpheme and cross-morpheme phonotactics, it is also consistent with cases of mismatch because reanalysis occurs only when there is uncertainty in the morphophonology.

CHAPTER 2

Modeling reanalysis

In this section, I outline a model of reanalysis that will be used to quantitatively demonstrate the effects of markedness biases in three empirical case studies. The purpose of modeling is not to generate novel results, but to help us understand the interaction of different variables in a system (Nigg 1994). More concretely, modeling allows us to probe at phenomena that are not directly observable, and also explicitly test intuitions about how certain patterns have arisen. In the current project, the goal of modeling is to understand how different variables affect reanalysis.

In Chapters 3-5, I present case studies where reanalysis is argued to be best explained by the interaction of frequency-matching with a markedness bias. However, it is also important to consider other factors such as types of frequency (type vs. token) and different learning biases (e.g. perceptual similarity bias). Computational models allow us to do just this, by quantifying and comparing the effects of different factors that may contribute to reanalysis in one direction or the other. In the context of language change (and specifically reanalysis), modeling is a particularly helpful tool. This is because reanalysis happens over generations of speakers, and is often inferred from limited historical data. Fine-grained data on all the intermediate stages of a change are rarely available, making it hard (or impossible) to find direct evidence for a hypothesis.

The model that I adopt is based in Maximum Entropy Harmonic Grammar (MaxEnt; Smolensky 1986; Goldwater & Johnson 2003), a probabilistic variant of Optimality Theory. Learning biases are implemented as a Gaussian prior, following the methodology laid out by Wilson (2006) and White (2013, 2017). Finally, to mirror the cumulative

effect of reanalyses over time, the model has an iterative (generational) component, in which the output of one iteration of the model becomes the input for the next. In the rest of this chapter, I will go over each component of the model.

(10) Toy data: final devoicing

STEM	SUFFIXED (PL)	UR	
bet	bed-a	/bed/	'cat'
mot	mot-a	/mot/	'dog'

For illustrative purposes, throughout this chapter I will use a toy example taken from Pater et al. (2012). In this toy language, exemplified in (10), stops are always voiceless word-finally and may alternate in voicing intervocalically, in this case when suffixed with the plural /-a/. The classic analysis for a pattern like this is that voicing is contrastive, but neutralized word-finally due to a process of final devoicing. Therefore, the word for 'cat' has the UR /bed/, where /d/ is devoiced to [t] in the bare stem. In contrast, 'cat' ends in an underlying /t/, so the final consonant is non-alternating.

If a speaker of this language is presented with a stem like [pat] and they have never heard the suffixed form, they must decide if the final [t] will be non-alternating (corresponding to UR /pat/) or alternating (corresponding to a UR /pad/). This ambiguity, illustrated in (11), can result in the learner reanalyzing a stem. For example, suppose that the original plural form of [pat] was [pat-a] (with a UR /pat/). If the language-learning child mis-learned the suffixed form to be [pad-a], and this change was passed down to the next generation of speakers, it would represent reanalysis in the direction of t→d.

(11) *Ambiguity in final obstruents*

$$[pat] \begin{cases} \nearrow [pad\text{-}a] \ /pad/ \\ \searrow [pat\text{-}a] \ /pat/ \end{cases}$$
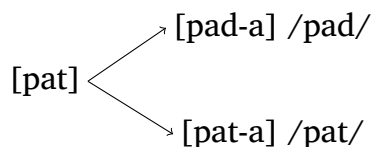
Existing probabilistic models of reanalysis predict these changes to be in the direction of the most likely alternant (see Chapter 1 for a review). Suppose that the toy lexicon had

the distribution given in (12), where 70% of final [t]'s are non-alternating. A statistical learning model would predict reanalysis to be mostly in the direction of d→t. On the other hand, if reanalysis happens in the direction of t→d, something other than statistical learning would be needed to explain the data. In this example, one possible cause could be a markedness bias against intervocalic voiceless consonants; such a bias would disfavor outputs like [mota] with an intervocalic [t], but not [beda]. The model I adopt must be able to capture frequency-matching behavior, while also accounting for effects of different learning biases.

(12) *Example: distribution of alternants*

TYPE	EXAMPLE	N
t∼d	[bet]∼[beda]	30
t∼t	[mot]∼[mota]	70

2.1 A MaxEnt model of reanalysis

Because I am looking at gradient (as opposed to categorical) alternations, I adopt Maximum Entropy Harmonic Grammar (MaxEnt; Goldwater & Johnson 2003; Smolensky 1986), a probabilistic variant of OT which uses weighted (instead of ranked) constraints, and generates a probability distribution over the set of candidate outputs.

In principle, other stochastic models of morphophonological learning may also work as models of reanalysis. For example, the Minimal Generalization Learner (Albright & Hayes 2003), introduced in Chapter 1, uses probabilistic rules, rather than constraints, to encode morphophonological alternations.

I adopt a constraint-based approach for the following reasons. First, I am looking at markedness effects in morphophonology, which I also argue to be restricted by stem phonotactics. OT straightforwardly captures product-oriented generalizations of this type, and more importantly can enforce phonotactics and alternations using the same con-

straints. In other words, phonological effects in morphology can be captured by an output-optimizing approach, where markedness is ranked above faithfulness (McCarthy & Prince 1993). For example, in our toy example above, a bias against intervocalic stops might be formalized in terms of a constraint *VTV, which penalizes both bare stems like [pata] and suffixed forms like [mot-a].[1]

In contrast, rule-based accounts (SPE; Chomsky & Halle 1968) are source-oriented, meaning that are described in terms of the input. While a bias against VTV sequences can be enforced using a rule of intervocalic voicing (t→d/V_V), there is no linking mechanism between stem phonotactics and alternations. In fact, in rule-based generative phonology, alternations are enforced using regular phonological rules, while stem phonotactics are enforced using context-free Morpheme Structure Rules/Constraints prior to the application of phonological rules (Halle 1959; Stanley 1967; Chomsky & Halle 1968). The two are treated as separate even though they often achieve the same goal (Duplication Problem; Kisseberth 1970; Kenstowicz & Kisseberth 1977).

Work by Bybee and colleagues (Bybee & Slobin 1982; Bybee & Moder 1983; Bybee 2003) suggests that speakers form both source-oriented and product-oriented generalizations. Albright & Hayes (2003) similarly find that while most speaker generalizations about English past tense formation can be formalized as input-output mappings , a subset are better characterized as generalizations on output forms. One potential weakness of my current approach is that compared to rule-based approaches, constraint-based frameworks are less suited to capturing source-oriented generalizations. Nevertheless, as noted by Baković (2007), constraints can be formalized to capture the rule-based generalizations.

---

[1]As a caveat, Paster (2005, 2009) finds that some morphophonological patterns cannot be captured using this output-optimizing approach. In particular, phonologically conditioned suppletive allomorphy requires additional mechanisms such as subcategorization constraints.

2.1.1 MaxEnt learning algorithm

In MaxEnt grammars (Goldwater & Johnson 2003), constraints are not ranked, but are instead weighted. Each input-output pair $[x_i, y_{ij}]$ is assigned a harmony (H_{ij}), which is the weighted sum of its constraint violations. Harmony is calculated using the equation in (2.1), where m is the number of constraints, w_m is the vector of constraint weights, and f_m is the vector of constraint violations.

$$H_{ij} = \sum_m w_m f_m(x_i, y_{ij}) \tag{2.1}$$

Harmony is more broadly a property of the family of Harmonic Grammars which use weighted constraints (Smolensky & Legendre 2006). In MaxEnt, harmony is mapped onto probabilities, where $p(y_{ij}|x_i)$ (the probability of an output y_{ij} given an input x$_i$) is calculated by taking the negative exponential of the harmony and normalizing this value by input as in (2.2).

$$p(y_{ij|x_i}) = \frac{1}{Z_i} e^{H_{ij}}$$
$$Z_i = \sum_{j'} e^{H_{ij'}} \tag{2.2}$$

The tableau in (13) illustrates how constraint evaluation works using our toy language. Final devoicing is enforced using a constraint NOFINALVOICE. Because NOFINALVOICE strongly outweighs the competing faithfulness constraint IDENT[voice], voiced stops are devoiced word-finally. This is shown for the input /bed/; candidate (a) is faithful but violates NOFINALVOICE, and is therefore assigned a higher harmony score. This in turn translates to a lower probability (P≈0). For ease of reading, predicted probabilities under 10^{-3} will be written as 0 in all following tableaux.

When /bed/ is suffixed, the faithful candidate (c) [beda] incurs no markedness violations. In contrast, candidate (d), which undergoes a voicing alternation that actually in-

creases violations of *VTV, is assigned a higher harmony and correspondingly a near-zero probability. Finally, given an input such as /mot-a/, the faithful candidate (e) violates a markedness constraint *VTV, which penalizes voiceless intervocalic stops. However, because *VTV has zero weight, it does not affect harmony, and candidate (e) still has a predicted probably of around 1.

(13) *Tableau: final devoicing*

	Obs	NoFinalVoice 12	Ident[voice] 6	*VTV 0	$\mathcal{H}$	e^{-H}	Z	$P = \dfrac{e^{-H}}{Z}$
/bed/								
a. [bed]	0	1			12	6.1×10^{-3}	2.5×10^{-3}	$3.35 \times 10^{-4} \approx 0$
b. [bet]	1		1		6	2.5×10^{-3}	2.5×10^{-3}	$0.999 \approx 1$
/bed/ + /a/								
c. [beda]	1				0	≈ 1	≈ 1	≈ 1
d. [beta]	0		1	1	6	2.5×10^{-3}	≈ 1	≈ 0
/mot/ + /a/								
e. [mota]	1			1	0	≈ 1	≈ 1	≈ 1
f. [moda]	0		1		6	2.5×10^{-3}	≈ 1	≈ 0

Unlike classical OT, where strict ranking ensures that losing candidates never surface, all candidates in MaxEnt grammars receive some probability. However, if constraint weights are sufficiently different, MaxEnt produces results that are functionally very similar to classical OT, where the winning candidate gets near-perfect probability. In fact, Johnson (2002) shows that as there is a finite limit on the number of constraint violations, there is a corresponding MaxEnt analysis for any classical OT analysis (of categorical data). When constraint weights are similar to each other, the model will predict variation and assign multiple candidates non-negligible probabilities.

The tableau in (13) uses hand-fitted weights, but in practice the learning problem of MaxEnt is to find weights that maximize the probability of the observed data $Pr(D)$. This value, also known as *likelihood*, is essentially the joint probability of all outputs y_{ij} given their respective inputs x_i. The equation for calculating likelihood is given in (2.3), where $p(y_{ij}|x_i)$ is the conditional probability of each output candidate y_{ij} given its input x_i, while *n* is the observed frequency of each input-output pair.

$$Pr(D) = \prod_i \prod_j p(y_{ij}|x_i)^n \qquad (2.3)$$

Because probabilities are being multiplied in (2.3), $Pr(D)$ will become extremely small as the number of possible output forms increase. In practice, it is therefore computationally easier to optimize the *log likelihood* given in (2.4), which is the sum of the logs of each $p(y_{ij}|x_i)$. The log function is monotonic, so minimizing log likelihood achieves the same result as minimizing likelihood.

$$\log(Pr(D)) = \sum_i \sum_j n \log(p(y_{ij}|x_i)) \qquad (2.4)$$

To summarize, MaxEnt optimizes an *objective function*, which in this case is the log likelihood of the observed outputs given in (2.4). The search space of log likelihoods is convex and therefore the optimal set of weights can be found by any standard optimization algorithm (Berger et al. 1996). The resulting model is will match rates of alternation in the lexicon and predict *frequency-matching* behavior.

In this dissertation, constraint weights were learned using the R package *maxent.ot* (Mayer et al. 2022), which uses the Limited-memory BFGS optimization algorithm (Malouf 2002) implemented in the *optim* function from the R-core statistics library. Constraint weights are also restricted to finite, non-negative values.[2]

---

[2]Nearly identical results are found using other gradient-based optimization methods such as the Excel Solver (Fylstra et al. 1998), which uses the Conjugate Gradient Descent method.

2.1.2 Learning biases in MaxEnt

In addition to log likelihood, the objective function in MaxEnt can also include a regularizing bias term, often referred to as a *Gaussian prior*. The prior is defined over each constraint weight, defined in terms of a mean μ and standard deviation σ[3]:

$$prior = \sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2} \tag{2.5}$$

In a model which takes into account the prior, the objective function is the log likelihood *subtracted* by the prior, as given in (2.6). The goal of learning is now to both maximize the log likelihood and minimize the prior.

$$\text{objective function} = \sum_i \sum_j n \log(p(y_{ij}|x_i)) - \sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma_i^2} \tag{2.6}$$

The numerator of this prior term is the squared difference of each constraint weight and its associated μ value. Consequently, as a constraint weight deviates from its μ, the penalty imposed by the prior increases. We can therefore think of the μ for each weight as its *a priori* preferred weight.

The other parameter, σ, determines how strongly each constraint weight is tied to its μ. When σ is large, the denominator will be large, meaning that deviations from μ will only incur a small penalty. In contrast, when σ is small, a greater penalty will be incurred when weights deviate from their μ. In other words, the smaller σ is, the more data is required to move weights away from μ during learning.

When the prior is uniform (i.e. μ and σ are the same for all constraints), the model prefers grammars where weight is even distributed among constraints. For this reason, Gaussian priors are often used in MaxEnt models as a way to prevent overfitting (e.g. Goldwater & Johnson 2003; Martin 2011).

---

[3]The prior is Gaussian in the sense that when converted to a probability space, it is a Gaussian distribution with mean μ and standard deviation σ

Following the approach developed by Wilson (2006), I use a non-uniform prior to implement bias. In principle, bias can be implemented by varying either σ (e.g. Wilson 2006) or varying μ (e.g. White 2013, 2017). I follow White (2013) and implement bias by varying μ values while keeping σ at a constant, relatively low value. More concretely, constraints with high μ values will prefer to have high weight, while those with low μ will prefer to have lower weight. This is demonstrated for our toy devoicing example in (14). The tableau is mostly identical to (13), with the added difference that each constraint is now associated with a μ and σ. Crucially, *VTV is assigned a higher μ value than competing faithfulness constraints. As a result, the model is biased to learn a higher weight for *VTV and ends up slightly dispreferring candidates like (e) (/mot-a/ →[mota]), which violate *VTV.

(14)   *Tableau: final devoicing with a markedness bias*

		NoFinVoice	Ident[voice]	*VTV	𝓗	P
	w	90	4.5	0.91		
	μ	90	0	1		
	σ	1	1	1	𝓗	P
/bed/						
a. [bed]	0	1			9	0
b. [bet]	70		1		4.5	1
/bed/＋/a/						
c. [beda]	30				0	1
d. [beta]	0		1	1	5.4	0
/mot/＋/a/						
e. [mota]	70			1	0.9	0.97
f. [moda]	0		1		4.5	0.03

This method of implementing bias as a prior term predicts that as the amount of input data increases, learners become less sensitive to learning biases. This is because

the objective function has two terms, the log likelihood and the prior. Log likelihood will increase as the number of observed input-output pairs increase, while the prior term will not (since it is a function of constraint weights rather than inputs). All else staying equal, this means that log likelihood will become more influential relative to the prior as the amount of training data increases.

I argue that this is a desirable consequence in line with findings from acquisition. Many theories of acquisition predict that markedness plays a stronger role in early learning, while frequency-matching becomes more prominent as the learner's lexicon increases; examples include the Tolerance Principle (Yang 2016) and Frequency Hypothesis (Levelt & Van de Vijver 1998/2004; Levelt et al. 2000). Various studies have also found empirical support for this learning trajectory (e.g. Levelt et al. 2000; Gnanadesikan 2004; Jarosz 2010).

In this dissertation, I look specifically at how phonotactic markedness effects can influence the learning of paradigms. A large body of work shows that children acquire phonotactics earlier than morphophonological alternations, and are moreover able to acquire fine-grained statistical generalizations about the phonotactics from an early age (for a review, see Sundara et al. 2022). It is therefore plausible that learners rely more on phonotactics in the early stages of paradigm learning. As their evidence for a morphophonological paradigm increases, they become better frequency-matchers.

2.2   Reanalysis as UR inference

In morphophonemic learning, learners must acquire both underlying representations and a phonological grammar mapping from these URs to surface representations. In other words, learners must concurrently learn Input-UR-SR mappings, where the Input is something like meaning or intent, typically encoded as morphosyntactic features. Various work has tackled this problem, including: models that use ranked constraints with error-driven learning (e.g. Tesar et al. 2003; Apoussidou 2006; Merchant 2008), maximum

likelihood learning in probabilistic frameworks (e.g. Jarosz 2006; Pater et al. 2012; Nelson 2019; Tan 2022), distributional learning with metrics such as Minimum Destription Length constraints (Rasin & Katzir 2016), rule-based frameworks (e.g. Rasin & Katzir 2020), and Bayesian Program Synthesis with ordered rules (Ellis et al. 2022).

The issue at hand is slightly different: when learners already possess a morphophonological grammar, they must be able to use it to infer URs from incomplete paradigms. Reanalysis happens precisely because learners, when given incomplete paradigms, sometimes infer a UR that differs from that of the previous generation of speakers. In particular, such changes enter the lexicon and result in language change, which is only possible if innovative URs can be inferred and listed in the lexicon. Results from wug testing also show that people are able to synthesize novel forms from incomplete data (e.g. Zuraw 2000; Ernestus & Baayen 2003; Hayes et al. 2009; Zuraw 2010a; Becker et al. 2011; Kawahara 2012; Gouskova & Becker 2013, and many more).

Existing models of Input-UR-SR learning assume that learners have access to complete paradigms, and therefore do not straightforwardly extend to this task. Where learning models have been tested against wug data or used to predict reanalysis (e.g. Albright & Hayes 2003; Ernestus & Baayen 2003; Calderone et al. 2021), they do not make use of URs.

In this section, I outline a procedure for UR inference, then discuss a few alternatives, each rooted in different assumptions about whether and how learners infer URs for stems with incomplete paradigms. In the context of the work described in this dissertation, the choice among these models is not essential; all alternatives can be used to model the interaction of frequency effects with a markedness bias.

2.2.1 UR inference

Again, I make the assumption that learners have already acquired a morphophonological grammar which includes a lexicon of URs, a procedure for building URs from complete

28

paradigms, and a phonological grammar mapping from URs to SRs. The main task of the learner in reanalysis is instead to infer URs when faced with structural ambiguity. For example, suppose that a speaker of the toy devoicing language would like to produce the plural form of [pat] 'rabbit', but they have never heard it before. They must then decide whether rabbit has a UR /pat/ or /pad/.

I suggest that this UR inference system is only deployed as needed, when the speaker has to infer a missing paradigm slot for a given input. When learners have access to a complete paradigm (or enough entries in a paradigm to resolve any structural ambiguity), they instead call on their morphophonological grammar to build URs.[4]

In many ways, an UR inference system resembles typical OT accounts, with key differences in the choice of input and output candidates. In UR inference, the input is not a phonological UR, but rather something like a lexical entry that includes meaning (encoded as morphosyntactic features) and associated phonetic realizations. This idea of an input that has no phonological material has been explored in various prior work, including Russell (1995a), Boersma (1998), Zuraw (2000), Wolf (2008), Pater et al. (2012), and Smith (2015). For example, to derive the suffixed form of 'rabbit', the input would be |RABBIT, [pat]| + |PL, [a]|. Note that I give inputs in vertical brackets (e.g. |CAT, [bet]|), URs in slash brackets (e.g. /bed/) and derived SRs in square brackets (e.g. [bet]). Additionally, where it is not relevant, I omit phonetic realizations from the input (e.g. |RABBIT + PL|).

The candidate set consists of UR-SR pairs, where UR→SR mappings are already decided by the learner's existing phonological grammar. The table in (15) shows examples of candidate URs inferred from the input |RABBIT + PL|. I specifically show UR inference for the *plural suffixed* form because again, by assumption, UR inference is only active when speakers are resolving structural ambiguity in a specific paradigm slot. In our toy language, this means suffixing environments which force speakers to "undo" word-final voicing neutralization.

---

[4]In §2.2.5, I discuss how speakers might resolve competing URs in the case where they infer a UR but subsequently receive inputs which contradict this UR.

For |RABBIT + PL|, two obvious candidate URs are /pat-a/ and /pad-a/ (see Section 2.2.2 for further discussion of the candidate set). The right-hand column of (15) shows the SRs derived from each UR. In cases where words show no within-item variation, each UR has only one corresponding SR determined by the phonological grammar (i.e. UR→SR mapping). For example, given a UR like /pad-a/, the SR [pata] is ruled out because in the lexicon, underlying /d/ never devoices to [t] intervocalically (e.g. /bed/ 'cat' is never observed with suffixed form [beta], which is presumably ruled out by faithfulness constraints).

(15) *Possible derivations of RABBIT-SUFF*

UR inference	Phon. grammar
Input→UR	UR→SR
|RABBIT-PL|→/pat-a/	[pata]
|RABBIT-PL|→/pad-a/	[pada]

Much like a typical phonological grammar, UR inference can be modeled in MaxEnt. To do this, I first introduce UR Inference constraints, defined in (16), which require surface realizations to mapped be to a particular UR. These UR Inference constraints can be context sensitive. For example, a constraint specifying that *final* [t] should be underlying /t/ can be written as "[t] = /t/, _#", or in shorthand as [t#] = /t/.

(16) *UR Inference constraint*

[a]→/b/, C_D: Assign one violation for every phonetic realization [a] in the input that does not correspond to UR /b/ in context C_D.

This constraint formulation is very similar to UR constraints, which have been proposed in work such as Zuraw (2000) and Pater et al. (2012). The key difference is that UR constraints are lexically specific mappings of *word meaning* to URs (e.g. |CAT|→/bed/), rather than mappings of surface phonetic realizations to URs (e.g. [t]→/d/). UR inference constraints also resemble *cue constraints* (Boersma 2007; Boersma & Hamann 2009;

Apoussidou 2006). However, cue constraints formalize the relation between auditory forms and phonological surface forms, while UR inference constraints map between surface forms and phonological underlying representations.

The model is trained on SR→UR pairs which parallel the environment where learners have to infer URs. In our toy examples, speakers have to infer URs when deriving a suffixed form from an unsuffixed form. The input is therefore UR-SR pairs such as [mot] + [a]~/mot-a/.

For example, we can assume the simplified lexicon in (17), consisting of lexical items with complete stem/suffix paradigms, input into the model as SR-UR pairs. 10 words are potentially alternating; of these, 7 words are underlyingly /t/-final, while 3 words are underlyingly /d/-final. The lexicon also contains words with final consonants that never alternate; for example, final [m] is always /m/ in the underlying representation. Finally, there are also words showing that medial [t] always corresponds to /t/.

(17) *Toy lexicon*

Type	SR	UR	N
final [t]→/t/	[mot] + [a]	/mot-a/	7
	[pit] + [a]	/pit-a/	
	...		
final [t]→/d/	[bet] + [a]	/bed-a/	3
	[hat] + [a]	/had-a/	
	...		
[m] is always /m/	[kom] + [a]	/kom-a/	5
	...		
medial [t] is always /t/	[patar] + [a]	/patar-a/	10
	...		

This training data can be evaluated on UR inference constraints as demonstrated in tableau (18). It is essential under this approach that candidates are not just URs, but rather URs paired with SRs, where the SR is the one derived by each UR (based on the learner's phonological grammar). This reflects the empirical results of the current dissertation, where reanalysis is found to be sensitive to markedness effects. In other words, learners appear to consider the well-formedness of the derived *surface* form when inferring new URs.

Tableau (18) contains both general UR inference constraints (e.g. [t]=/t/) and positional ones (e.g. [t#]=/t/). The model learns a relative weighting of UR inference constraints that predicts frequency-matching behavior. First, [t]=/t/ has a much higher weight than [t]=/d/; this rules out candidates like (h) and (i), ensuring that medial [t] always corresponds to /t/. The context-specific constraint [t#]=/d/ has some weight, but is still smaller than [t]=/d/. This ensures that final [t] will sometimes be underlying /d/, but is still /t/ 70% of the time. Additionally, [t]=/t/ and [t#]=/d/ gang up to rule out unobserved SR-UR mappings such as [t]→[m] in candidate (c).

A UR inference model can also include markedness constraints that evaluate the derived SR corresponding to each UR. Tableau (18) includes one markedness constraint *VTV.[5] The model learns zero weight for *VTV, but a markedness bias can be implemented so that the model will preferentially learn a higher weight for *VTV.

---

[5]As discussed in Chapter 1, I argue that markedness constraints are restricted to active markedness effects already present in stem phonotactics. Additionally, subsequent chapters show how such constraints could be learned directly from a phonotactic model.

(18)   *UR inference model*

	Obs	[t#] = /t/	[t#] = /d/	[t] = /t/	[t] = /d/	[m] = /m/	*VTV	$\mathcal{H}$	P
		0	4.3	5.5	0.3	5.2	0		
[CVt] + [a]									
a. /CVt-a/ [CVta]	7		1		1		1	4.6	0.70
b. /CVd-a/ [CVda]	3	1		1				5.5	0.30
c. /CVm-a/ [CVma]	0	1	1	1	1			10.1	0.00
[CVm] + [a]									
d. /CVm-a/ [CVma]	5							0.00	0.99
e. /CVd-a/ [CVda]	0					1		5.2	0.01
f. /CVt-a/ [CVta]	0					1	1	5.2	0.01
[VtV] + [a]									
g. /VtV-a/ [VtV-a]	7				1		1	0.3	0.99
h. /VdV-a/ [VdV-a]	0			1				5.5	0.01
i. /VmV-a/ [VmV-a]	0			1	1			5.8	0.00

This is demonstrated in (19), which shows predictions of a model that was run with the same inputs and constraint set as (18), but has the addition of a prior term. To implement a bias against intervocalic [t], *VTV can be assigned a higher μ than competing constraints. In this model, $\mu = 3$ for *VTV, $\mu = 0$ for all other constraints, and $\sigma = 5$ for all constraints.

The resulting model is similar to the one in (18), except that it learns a non-zero weight (w = 1.4) for *VTV. As a result, the model slightly over-predicts rates of [t]∼[d] alternation (P = 0.33 vs. P = 0.30). Inclusion of a bias term can therefore account for UR inference that is *not* frequency-matching. Additionally, even though the inclusion of a bias results in just a small difference in predicted probability of different candidates (3% in this example), such changes can accumulate over generations of reanalysis; I discuss how such cumulative effects are modeled in the following section.

(19) *UR inference model with a bias*

		$\|t\#\| = /t/$	$\|t\#\| = /d/$	$\|t\| = /t/$	$\|t\| = /d/$	$\|m\| = /m/$	*VTV		
	μ	0	0	0	0	0	3		
	w	0.6	4.3	5.8	0	4.7	1.4	$\mathcal{H}$	P
[CVt] + [a]									
a. /CVt-a/ [CVta]	7		1		1		1	5.6	0.67
b. /CVd-a/ [CVda]	3	1		1				6.4	0.33
c. /CVm-a/ [CVma]	0	1	1	1	1			10.7	0.00

The resulting trained model can then be used to infer URs for words with incomplete paradigms. To form the UR for |RABBIT + PL|, the learner samples from possible candidates across probability distribution. This means that, given the grammar in (19), they have a ≈67% chance of selecting /pat/ to be the UR.

2.2.2 Candidate set (Input URs)

In the tableau given so far, I assume a relatively small candidate set of URs. The candidate set of URs could potentially be much larger. For example, given the surface form [pat], possible URs could include /pat/ and /pad/, but also candidates like /pak/ (which diverges further from the observed surface form) and /paD/ (where D is an abstract phoneme that is never realized faithfully in the surface form). This notion of UR abstractness is formalized by Kenstowicz & Kisseberth (1977, Ch. 1), who set up a taxonomy of UR-SR distance.

Deciding the level of UR abstractness—or how much URs can diverge from their SRs—is not central to the UR inference system. This is because I assume that learners have acquired a morphophonemic grammar that includes mappings of URs to their allophonic variants. As shown above in tableau (18), SR-UR mappings such as [t] = /m/ are ruled out

by the UR inference grammar, in the same way that unobserved candidates in a traditional phonological grammar are eliminated via constraint weighting or ranking.

Nevertheless, it is worth noting that increasingly, findings from language change and learning experiments support a more restrictive view of URs, where URs are constrained to be closer to their surface realizations. For example, Kiparsky (1965, 1982) finds evidence that patterns which are amendable to abstract UR analyses tend to be unstable, and are often removed by subsequent language change. Recent experimental work supports this view, showing that such "abstractness-friendly" patterns may be learnable, but are harder to learn (e.g. White 2017).

2.2.3   What is the base of UR inference?

In the example discussed so far, I assume that only one paradigm slot (the bare stem) serves as the base of UR inference. In principle, however, any paradigm slot available to the learner can be a base for UR inference.

In a situation where multiple possible paradigm slots are available to the speaker, there are several possibilities for how speakers choose between them. Speakers may preferentially select the bare stem due to its privileged unmarked status (e.g. Kuryłowicz & Winters 1947; Mańczak 1957). They might also select the member of the paradigm with the highest token frequency (Mańczak 1980, p. 284-285).

Following Albright (Albright 2002a,b, 2010, etc.), I propose that learners select the (available) member of the paradigm that is most *informative*. A paradigm slot is informative in the sense of being able to predict other paradigm slots with high accuracy, on the basis of statistical regularities. Where available paradigm slots are equally informative, other factors (such as frequency and markedness) may then come into play; this idea is discussed in Albright (2008).

2.2.4   Probabilistic UR-SR mappings

Throughout the examples discussed so far, I have assumed that each UR has just one corresponding SR, so learning probabilities over observed surface forms is equivalent to learning probabilities over URs.

In tableau (9), for example, each UR has just one corresponding SR. This is true for the devoicing example and more generally for the case studies I consider in the following chapters. However, languages often have gradience that causes the same underlying sound (in the same context) to have variable surface realizations (e.g. see Zuraw & Hayes 2017 for three case studies). In other words, the phonological grammar generates a probability distribution over SRs for each UR.

When there are variable surface realizations for the same UR, learning becomes more complex. For example, supposed that a in our toy language, underlying intervocalic /d/ sometimes surfaces as [t]. Model inputs would look like in (20); this tableau is the same as (19) except for the addition of candidate (b), where underlying /d/ is realized as [t]. This candidate violates standard IO-faithfulness constraints such as IDENT-IO[voice], and also incurs violations of the relevant UR Inference constraints.

(20)   *UR inference model: multiple SRs*

[CVt] + [a]	SR	Obs	|t#| = /t/	|t#| = /d/	|t| = /t/	|t| = /d/	IDENT[voice]	*VTV
a. /CVt-a/ [CVta]	[CVta]	7		1		1		1
b. /CVd-a/ [CVta]				1	1		1	1
c. /CVd-a/ [CVda]	[CVda]	3	1		1			
d. /CVm-a/ [CVma]	[CVma]	0	1	1	1	1		

The model still learns probabilities over observed SRs, but now these values are summed over multiple UR candidates. In particular, the SR [CVta] can correspond to either can-

didate (a) or candidate (b). The resulting hidden structure means that the search space is no longer convex and optimization is not guaranteed to converge on the optimal solution.

2.2.5 Sources of competing URs

UR inference occurs only when learners have incomplete access to a paradigm. Presumably, learners can then list the newly inferred UR and use it for deriving surface forms. However, this process implies that learners could end up with multiple *listed* URs.

For example, a leaner could infer a UR that differs from the one adopted by the general speech community. They may therefore receive subsequent input that resolves structural ambiguities in a paradigm, but conflicts with the listed UR. For example, a learner might infer a UR /pat/ for rabbit, but then hear the suffixed form [pada], which tells them that the UR is in fact /pad/. As a result, the learner's grammar now has two competing listed URs.

UR inference is also tied to the paradigm slot that the speaker is trying to produce, and to the markedness properties of the derived output form in that paradigm slot. As a result, model predictions may differ depending on the paradigm slot that a learner is trying to produce. For example, suppose that our toy language has two suffixes, respectively vowel-initial /-a/ 'PLURAL' and consonant-initial /-ka/ 'DIMINUTIVE'. If a speaker were trying to infer a UR for |RABBIT + PL|, the non-alternating candidate /pat-a/ [pata] would violate *VTV. If they were instead inferring the UR for |RABBIT + DIM|, the non-alternating candidate /pat-ka/ [patka] incurs no violations of *VTV.

The consequences of this contrast is illustrated in tableau (21). This current tableau shows model predictions for two inputs, |pat + a| and |pat + ka|. The model was trained on our toy lexicon, with a bias towards high weight for *VTV; constraint weights are taken from (19). Crucially, candidate (a) and (d) both have an underlying stem-final /t/, but (a) incurs a violation of *VTV while (d) does not. Because *VTV has non-zero weight, candidate (a) is assigned a lower predicted probability than candidate (d).

37

(21)   *Predicted URs across different suffixal environments*

	|t#| = /t/	|t#| = /d/	|t| = /t/	*VTV	$\mathcal{H}$	p
	0.59	4.28	5.8	1.38		
|RABBIT,[pat]| + |PL,[a]|						
a. /pat-a/ [pata]		1		1	5.66	0.67
b. /pad-a/ [pada]	1		1		6.39	0.32
c. /pam-a/ [pama]	1	1	1		10.67	0
|RABBIT,[pat]| + |DIM,[ka]|						
d. /pat-ka/ [patka]		1			4.28	0.89
e. /pad-ka/ [padka]	1		1		6.39	0.11
f. /pam-ka/ [pamka]	1	1	1		10.67	0

One way to account for competing listed URs is to encode each UR with a gradient 'memory strength'. This idea is explored in Moore-Cantwell & Pater (2016). Essentially, each UR is associated with a representational strength, which in MaxEnt can be encoded as the weight of a lexically-specific UR constraint (e.g. RABBIT = /pat/ and RABBIT = /pad/). The memory strength of each UR decays over time, but will increase if the learner encounters input data associated supporting this UR. For our 'rabbit' example, if the learner encounters the diminutive form of rabbit more than the plural, they might learn a stronger representation for /pat/ than /pad/.

2.2.6   Alternative approaches

The UR inference approach described in this section is able to capture statistical patterns within a paradigm, while also accounting for 'global' phonotactic markedness effects via a bias term. In essence, learners infer fully specified URs when undoing structural ambiguity in a paradigm. In this section, I discuss two alternative approaches, where URs are either underspecified or not posited for incomplete paradigms.

38

As previewed above, the choice of implementation is not central to my arguments; both alternatives are able to capture the interaction of frequency-matching and markedness bias. Nevertheless, I will argue that UR inference is a more accurate reflection of how reanalysis works, and that the two alternatives make undesirable predictions.

The first alternative to a UR inference analysis, formalized by Albright (2002b), is to assume surface representations. In other words, the model is trained on surface forms which serve as the base of reanalysis. The Minimal Generalization Learner (MGL Albright & Hayes 2003) is a rule-based implementation of this idea. Additionally, many models of wug test results (which, like reanalysis, addresses how learners 'fill' incomplete paradigms) do not make use of underlying representations, and therefore implicitly assume a surface-base account (e.g. Albright & Hayes 2003; Ernestus & Baayen 2003; Calderone et al. 2021).

In a constraint-based framework, the surface-base approach can be modeled by using a surface slot of the paradigm as the input; candidates are possible allomorphs of the paradigm slot that the learner is trying to derive. In our toy language, the input would be forms like [bet] 'cat' and [mot] 'dog', while candidate outputs are possible suffixed form allomorphs ([bet-a], [bed-a], [mot-a], [mod-a]).

This approach is illustrated in tableau (22). The tableau assumes a simplified 10-word lexicon where [t]~[d] alternation occurs in 3 forms (i.e. 30% of the time). Therefore, for [t] final input stems (written as [CVt]), the output candidate [CVt-a] has an observed frequency of 3, while [CVd-a] has an observed frequency of 7.

The input and candidate outputs are all *surface forms*. Faithfulness constraints therefore encode Output-Output relations (Benua 1995); in (22), for example, OO-IDENT[voice] requires the final consonant of the stem and suffixed forms to share the same voicing specification. The relative weights of faithfulness constraints and competing markedness constraints (e.g. *VTV) give rise to probabilistic rates of voicing alternation.

(22) *Predicting voicing alternations under a surface-base approach*

[CVt] + SUFF	Obs	*VTV 0.87	OO-ID[voice] 1.72	$\mathcal{H}$	P
a. CVt-a	7	1			0.701
b. CVd-a	3		1		0.299

The surface base grammar can be straightforwardly applied to stems where the suffixed form is unknown (and more generally to any novel items). This is demonstrated in (23) for the input [pat]. Assuming the learner does not have access to the suffixed form, they can derive it using the relative weighting of output-output faithfulness constraints and conflicting markedness constraints.

(23) *Predicting novel suffixed forms in the stem-base approach*

[pat] + SUFF	*VTV 0.87	OO-ID[voice] 1.72	$\mathcal{H}$	P
a. pat-a	1			0.701
b. pad-a		1		0.299

However, the surface-base approach makes no distinction between lexical forms and novel forms, and therefore predicts variation in both types of forms. The invariance of known lexical items can be resolved using various approaches to enforcing lexically-specific behavior (e.g. Zuraw 2000, 2010a; Pater 2007, 2008; Becker 2009). This solution would still treat reanalyzed forms as not having a UR, and does not resolve the issue of

how learners acquire a fully specified representation for reanalyzed forms.

The second alternative to UR inference is to underspecify structurally ambiguous parts of a stem. In other words, when learners are given a stem that is potentially alternating, they posit an underspecified UR that encodes only fixed aspects of the potential alternant. In the toy language, URs have fully specified final consonants when the learner has access to the suffixed form (/bed/ 'cat' and /mot/ 'dog'). On the other hand, for the word [pat] 'rabbit' whose suffixed form is unknown, the learner might posit the UR /paT/, where /T/ represents a coronal stop that is *not* specified for voicing.

In this way, words with complete paradigms get fully specified URs and are protected from variation by faithfulness constraints. For underspecified forms, the grammar (i.e. relative weighting of markedness constraints) can then fill in /T/'s voicing specification. This underspecification approach is taken up and fully fleshed out in work such as Inkelas et al. (1997).

The underspecification approach can capture statistical generalizations in reanalysis. However, it is fundamentally problematic because it characterizes reanalyzed forms as patterning differently from from other listed forms. Reanalyzed forms are underlyingly distinct from other listed forms (/T/ vs. /t, d/), which implies that have distinct phonological behavior. In fact, underspecification is generally used to capture ternary contrasts for this very reason (Inkelas et al. 1997).

2.3 Iterated learning

Models of language change are by no means new, and there are various approaches to doing so. Weinreich et al. (1968) use phonological rules that apply variably to predict change in progress. Other approaches that have been explored include modeling change in dynamical systems (Niyogi 2006), connectionist frameworks (Tabor 1994), as the result of competing grammars (Yang 2000), in exemplar-based frameworks (Pierrehumbert 2002), and more recently in variants of OT (e.g. Boersma 1998; Zuraw 2000, 2003).

To simulate the cumulative effects of reanalysis over time, I assume an agent-based iterated learning model. Under this approach, small changes to an alternation pattern can accumulate over iterations (each corresponding roughly to a generation of speakers), resulting in large-scale reanalyses of a pattern.

Generation 1   Generation 2   Generation 3   Generation n

H_1 *produce* $\longrightarrow$ *observe* H_2 *produce* $\longrightarrow$ *observe* H_3 *produce* $\cdots$ *observe* H_n
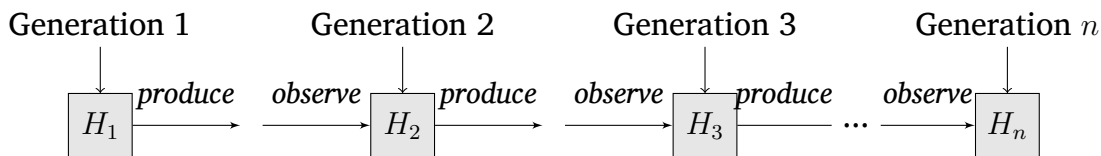
Figure 2.1: Structure of an iterated learning model, adapted from Ito & Feldman (2022, p. 3). H_i indicates hypotheses of each generation.

In an agent-based iterated model, the output of one model iteration becomes the input to the next iteration. The current study adopts a simplified model in which each generation (or iteration) has just one agent and one learner, as illustrated in Fig. 2.1. In the first generation, the agent A1 produces the output language based on their grammatical knowledge (i.e. Hypothesis 1; H_1). More concretely, a hypothesis is the speaker's grammar, represented in this case using MaxEnt, as the probabilistic weighting of Optimality-Theoretic constraints. The learner observes these data, induces the relevant generalizations and forms another hypothesis (H_2), which then becomes the basis of the output data presented to the next generation. This process is repeated for many iterations.

When providing input for a learner in the next generation, not all of the information of the language is presented, resulting in a learning "bottleneck" (Brighton 2002; Kirby 2001; Griffiths & Kalish 2007). As a result of this bottleneck, input patterns that are easier to learn should be more likely to pass through this bottleneck, and become more prominent over generations of learning. In the current study, the bottleneck is implemented by having the Agent "forget" some proportion of forms at each iteration. The remembered forms are retained to the next generation, while the forgotten forms are generated from the Agent's grammar (Hypothesis 1, 2, 3, etc.).

Most iterated learning studies have used artificial languages composed of very few items, and focused on explaining the emergence of broad characteristics of linguistic

structure such as compositionality (e.g. Kirby 2001; Brighton 2002; Griffiths & Kalish 2007). Where iterated learning has been applied to language-specific patterns of change over time, the focus is on predicting largely exceptionless sound changes. Studies that look at the restructuring of morphophonological paradigms are less common, and have almost all looked at the effect of (type vs. token) frequency distributions in driving re-analysis (e.g. Hare & Elman, 1995; Polinsky & van Everbroeck, 2003). As discussed in Chapter 1, these models predict that reanalysis will match distributions of the input data.

To the author's knowledge, Ito & Feldman (2022) is the only iterated learning study of morphophonological change that implements a bias against frequency-matching. Their study focuses on accent change in Sino-Korean, and specifically on how phonoactics inform analogical change. Like the current study, they find that models which include phonotactic information outperform purely distributional models.

Ito and Feldman's study differs from the current one in that they look at how a static pattern of word accentation is affected by segmental phonotactics. In contrast, my focus is on how static phonotactics can influence morphophonological alternations. Additionally, whereas I implement bias as a prior in MaxEnt, Ito and Feldman implement a sensitivity to phonotactics by using Bayesian inference.

Note that the iterated learning paradigm I adopt makes several simplifying assumptions. In particular, I assume just one agent and one learner, when in fact language change takes place at the level of the population. Future work should therefore consider more complex models which incorporate multiple interacting Agents in a way that models the speech community. Baker (2008) finds that such multi-agent models produce more empirically accurate results, and are also able to model sigmoidal progression of change.

I also do not consider how learning changes in adulthood. A more complex model may include an update algorithm for adult grammars, and also account for the stronger effect of markedness bias in early learning (when speakers have a relatively impoverished input).

2.3.1 Rate of change

The literature on lexical diffusion suggests that change often occurs non-linearly in an 'S-shaped curve' (i.e. sigmoidal function; Kroch 1989; Blythe & Croft 2012; Denison 2003; Hayes 2022). In the context of reanalysis involving two alternants A and B, this means that at first, alternant A is rarely used and its use increases slowly. Then, as A and B approach equal rates of alternation, alternant A will gain ground more and more quickly. Past this point of inflection, the rate of change slows down again.

My current approach predicts a more gradual trajectory for markedness-driven reanalyses. To understand why, consider first a baseline model with no markedness bias. In the iterated learning model setup, speakers fill in incomplete paradigms by sampling from the probability distribution generated by their MaxEnt model. Because MaxEnt is inherently frequency-matching, a model with no bias will, when averaged over multiple iterations, predict equal rates of each alternant over time. For a more concrete example, we can look at the devoicing example again; Fig. 2.2 shows the predicted proportions of each alternant (/t/ vs. /d/) over 50 iterations. Because random sampling introduces variation, values are the average of 30 model runs. The grey intervals indicate standard error, where a larger standard error indicates more variation between runs of the model.

In the left-hand unbiased model, we see that predicted rates of each alternant are generally stable over time (i.e. the model is frequency-matching). When a bias term is introduced, rates of alternation will change gradually, shifting by a roughly equal magnitude in each iteration. This is shown in the right-hand side of Fig. 2.2, which shows predicted results when the model is given a strong markedness bias against intervocalic stops.

Future work should therefore consider how to model rates of language change in a way that better match the S-shaped trajectory. One potential solution is to adopt more complex models which have multiple Agents, interacting in a way that models the speech community. For example, Baker (2008) finds that multi-agent models can predict a sigmoidal
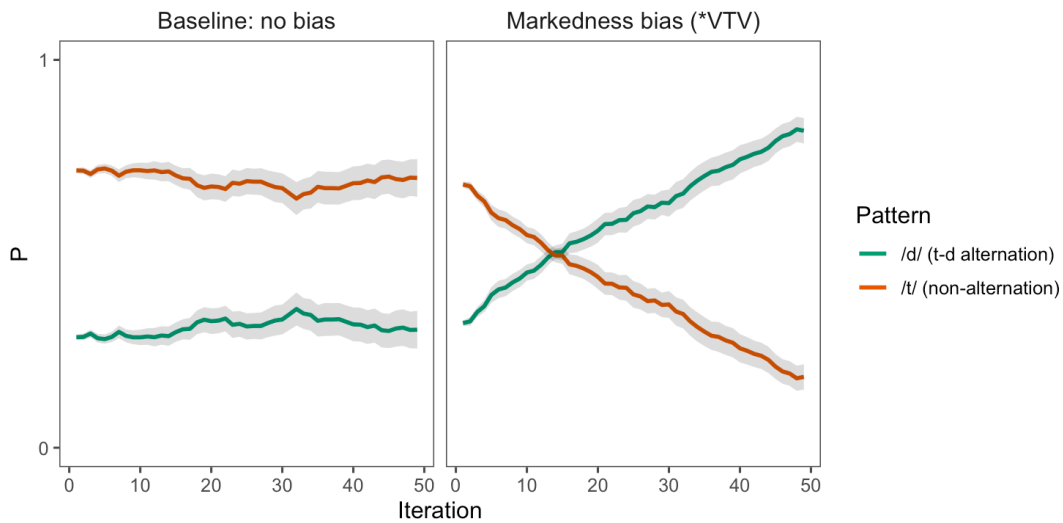
Figure 2.2: Effect of bias on predicted proportions of alternants across 50 iterations

progression of change if a speaker's probability of adopting the change is proportional to the prevalence of the change in the speech community.

2.3.2 Parameters in iterated learning

The iterative learning model has two parameters: forgetting rate and number of iterations. The forgetting rate is the proportion of forms forgotten and relearned in each iteration; in all subsequent studies, I test 5 forgetting rates (0.05, 0.1, 0.15, 0.2, 0.25).

In the simplest version of the model, words are sampled on a flat distribution so that every word has an equal probability of being forgotten. A more realistic model would factor in token frequency, such that high-frequency words are less likely to be forgotten. This builds on extensive work showing that words with high usage/token frequency resist reanalysis (e.g. Moder 1992; Bybee 1995; Polinsky & van Everbroeck 2003). In Chapter 3, I consider a model which accounts for token frequency.

In setting the number of iterations, I follow Ito & Feldman (2022) in equating each iteration to roughly one generation of speakers, where a generation lasts 25 years. Therefore, a change that occurred over around 100 years would take four iterations ($25 \times 4 = 100$).

Forgetting rate and the number of iterations are closely related; in general, when the forgetting rate is low, rate of change over time is slower, but this can be offset by increasing the number of iterations. However, increasing the forgetting rate has the additional effect of increasing variation between different runs of the model. This is because as forgetting rate increases, the input data for each model iteration becomes more variable. The effect of varying forgetting rate is demonstrated in Fig. 2.3, which shows predictions of the markedness-biased model when forgetting rate is varied; this figure uses the same devoicing example as Fig. 2.2 above. When the forgetting rate is very low, the model behaves similarly to the baseline model and is roughly frequency-matching. As forgetting rate increases, the strength of markedness effects in each iteration also increases. Notably, variation between model runs also increases, as evidenced by the higher standard error (i.e. wider grey interval in the figures).
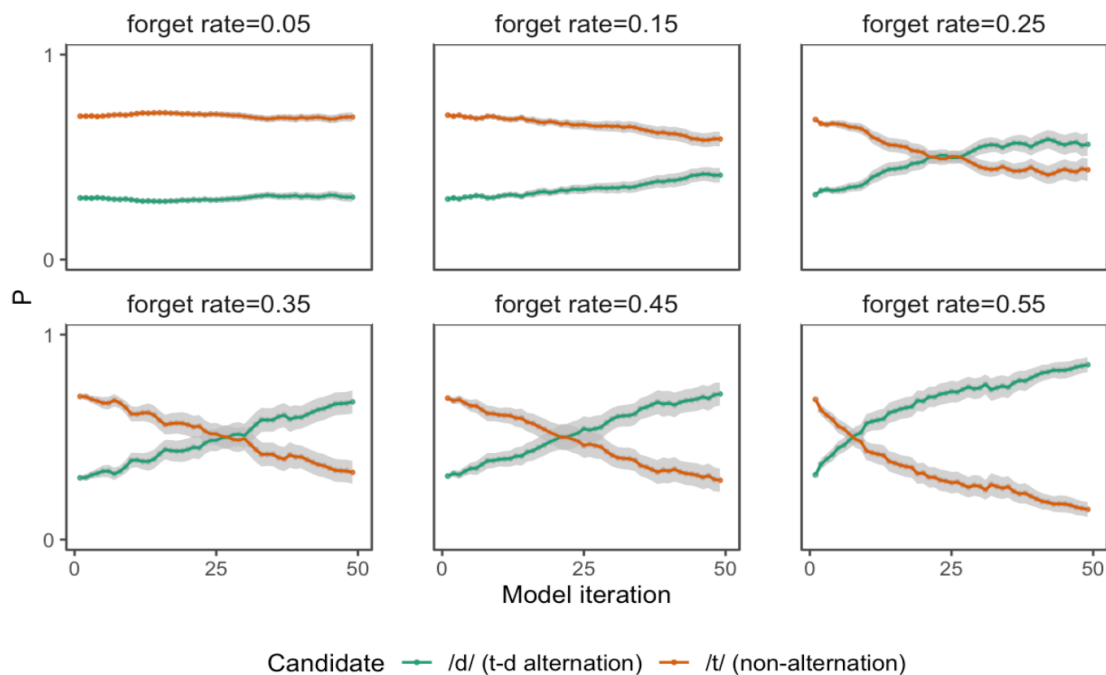
Figure 2.3: Predictions of a markedness-biased model over different forgetting rates

2.3.3 Iterative modeling and the choice of σ

In modeling reanalysis, superficially similar outcomes can be achieved by setting σ^2 to a lower value and reducing the number of iterations. A lower σ^2 allows the bias to have a stronger effect, so that the model predicts a greater magnitude of change per iteration. In other words, a low-sigma model can converge on nearly identical results as a high-sigma model within fewer iterations. Fig. 2.4 shows the model fit over 50 iterations for the markedness-biased model when σ^2 is varied and μ values are held constant. Both the high-sigma model ($\sigma^2 = 1$) and low-sigma model ($\sigma^2 = 0.1$) converge on similar outcomes, but the low-sigma model does so much faster, plateauing after less than 10 iterations.

σ = 0.1   σ = 1

Pattern
— /d/ (t-d alternation)
— /t/ (non-alternation)

P
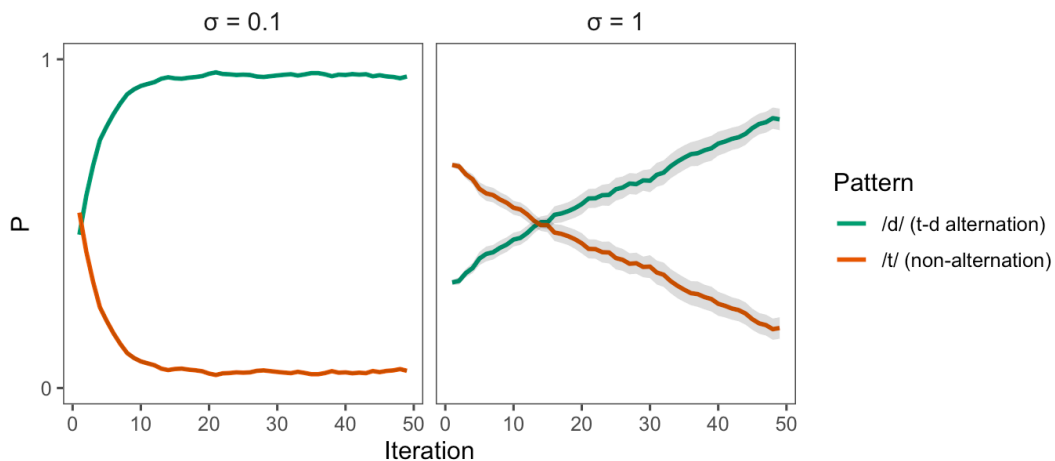
Iteration

Figure 2.4: Effect of varying sigma on predicted proportions of alternants

Although a low-sigma model achieves the same outcome as a high-sigma model, I argue that the high-sigma model, which takes more iterations to converge on the same result, is preferable for the following reasons. First, it is conceptually more plausible that reanalysis happens over many generations of speakers, instead of just 1-2 generations. A high-sigma model also predicts more randomness and variation. In my implementation, variation is introduced by randomly sampling the winning candidates that become the input to the next model iteration. In a high-sigma model, the effects of the bias are weaker, so there is more variation how winning candidates are sampled. In a low-sigma model, the markedness bias is so strong that the model will almost always select the

markedness-reducing candidate.

Fig. 2.5 compares 12 individual model runs of the markedness-biased when sigma is high ($\sigma = 1$, top figure), and when sigma is low ($\sigma = 0.1$, bottom figure). In the low-sigma model, there is almost no variation between model runs, and the model always converges on the same result after a few iterations. In the high-sigma model runs, reanalysis generally happens in the direction of the markedness-reducing alternant. However, in a subset of model runs (e.g. Runs 5 and 6), the model predicts a smaller magnitude of change. In some runs (e.g. Run 11), the model also predicts roughly equal rates of alternation for both alternants, with no clear shift towards the markedness-reducing alternant. This type of variation matches how language change happens in reality, where markedness bias may affect different languages to a different degree, and the same language will undergo dialect divergence.
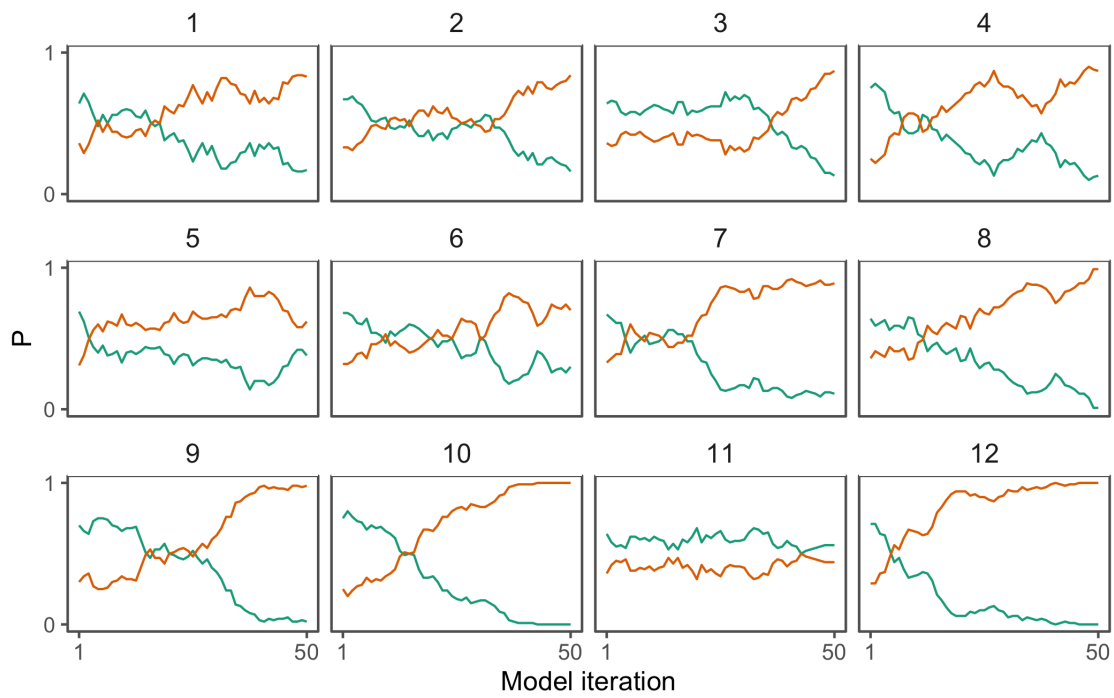
(a) High-sigma model ($\sigma = 1$)
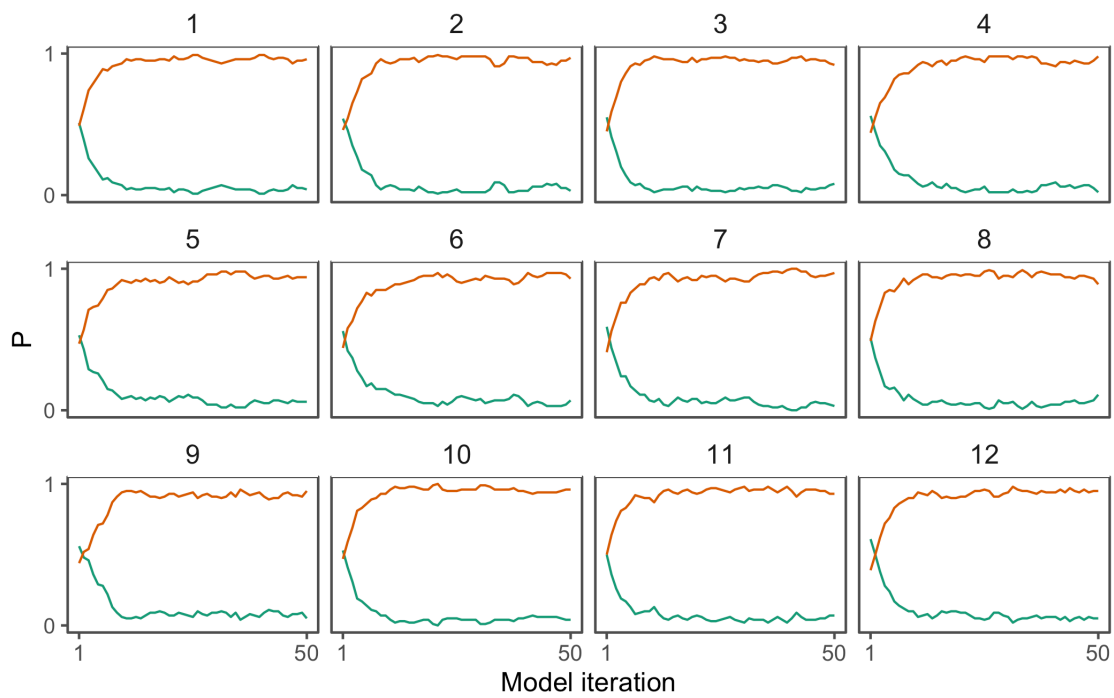
(b) Low-sigma model ($\sigma = 0.1$)

Figure 2.5: Individual model runs in low-sigma model (top) vs. high-sigma model (bottom)

CHAPTER 3

Case study 1: Malagasy weak stem alternations

In the this chapter, I present a case study of reanalysis from Malagasy, an Austronesian language spoken in Madagascar. Malagasy has inflectional and derivational morphology, much of which involves morphophonological alternations. In a subset of so called **weak stem** consonant alternations, the expected alternant (based on historical evidence) often does not match the observed alternant, suggesting that substantial reanalysis has occurred.

As a preview of the results, I find that for Malagasy, most weak stem alternations are predictable from statistical distributions within the paradigm. For one set of alternations, however, the direction of reanalysis contradicts the predictions of frequency-matching models. I propose that reanalysis in Malagasy can only be explained by a model which incorporates both frequency and markedness effects. Additionally, I will show that the Malagasy data is consistent with principles of active markedness.

The rest of the chapter is organized as follows: §3.1 presents background on Malagasy phonology and descriptive generalizations about the Malagasy weak stems. In §3.2, I present results of a corpus study comparing historical Malagasy forms with modern Malagasy data, to show that reanalysis has occurred in a way that cannot be predicted by purely frequency-matching models of reanalysis. In §3.3, I show that a model which incorporates a markedness bias outperforms frequency-matching alternatives. .

3.1 Background

Malagasy, the national language of Madagascar, is an Austronesian language belonging to the South East Barito subgroup of the Western Malayo-Polynesian subfamily (Rasoloson & Rubino 2005). The term Malagasy really refers to a macro-language that covers many dialects distributed throughout Madagascar (Lewis et al. 2014). This chapter uses data from Official Malagasy (OM), which is the standardized, institutional dialect that is based on the Merina dialect spoken in the capital city Antananarivo. All subsequent descriptions and analysis will assume data from OM.

Malagasy phonetics and (morpho)phonology is relatively well-documented. The phonetic system is described by Howe (2021). Additionally, I take descriptive generalizations of the morphology and phonology from Erwin (1996), Keenan & Polinsky (2017), and O'Neill (2015) (who documents the closely related Betsimisaraka dialect). Formal analyses of Malagasy phonology, including of weak stem alternations, have been done in both generative rule-based frameworks (Dziwirek 1989) and OT (Albro 2005). Additionally, as will be described below, I base statistical generalizations about the Malagasy lexicon off of data from the Malagasy Dictionary and Encyclopedia of Madagascar (MDEM; de La Beaujardière 2004), which compiles dictionary data from multiple Malagasy dictionaries.

Moreover, the history of Malagasy can be traced in some detail through the work of Austronesianists (e.g. Dahl 1951; Mahdi 1988; Adelaar 2013). Historical comparative data is also available in the Austronesian Comparative Dictionary (ACD; Blust & Trussel 2010).

In the rest of this section, I provide a descriptive account of Malagasy phonology and weak stem alternations, based on work by Keenan & Polinsky (2017) and Howe (2021).

3.1.1 Malagasy phonology

Malagasy words have a strict (C)V syllable structure, where codas are not allowed. Word stress is generally penultimate, with the following exceptions. A small set of disyllabic words have final stress; these are generally loans ([diˈte] 'tea' <French *du thé*) or native roots (especially demonstratives) ending in a diphthong [iˈzai] 'there/COMPARATIVE'. Among longer roots (3+ syllables), a subclass of words called 'weak stems' have antepenultimate stress; these will be discussed in the following section.

Malagasy has five phonemic monophthongs /i e a o u/. /o/ is considered to be non-phonemic (or marginally phonemic) in many descriptions of Malagasy (e.g. Rasoloson & Rubino 2005; O'Neill 2015). However, it has become much more common as a result of /ua/ and /au/ sequences merging to /o/ in OM (Howe 2021).

	bilabial	labiodental	dental	alveolar	retroflex	velar	glottal
plosives	p, b		t, d			k, g	
	^{m}p, ^{m}b		^{n}t, ^{n}d			^{ŋ}k, ^{ŋ}g	
affricates				ts, dz	ʈʂ, ɖʐ		
				^{n}ts, ^{n}dz	^{n}ʈʂ, ^{n}ɖʐ		
nasals	m		n			(ŋ)	
trills/flaps				r~ɾ			
fricatives		f, v		s z			h
lat. approximants				l			

Table 3.1: Malagasy consonant chart

The consonants of Malagasy are given in Table 3.1. /ŋ/ is given in parentheses because although it is non-phonemic in OM, it is phonemic in many dialects of Malagasy.

All subsequent examples are presented in IPA, with the following caveats. Prenasalized obstruents are written as nasal-obstruent sequences (e.g. *mb* corresponds to /^{m}b/). /ʈʂ/ and /ɖʐ/ are generally realized as retroflex, but can vary in production between speakers (Howe 2021), and have been described in prior work as post-alveolar (e.g.

Keenan & Polinsky 2017). In addition, [r] is a short alveolar trill in most dialects including OM, but is often realized as a tap [ɾ] in casual speech (Howe 2021). Additionally, medial [h] is pronounced in careful speech, but otherwise often elided. Unstressed vowels are also often devoiced or deleted (Howe 2021).

3.1.2 Weak stems

Malagasy has a class of forms that Keenan & Polinsky (2017) refer to as weak stems. Weak stems always end in one of the three 'weak syllables' /ʈʂa/, /ka/, or /na/.[1] Additionally, they have antepenultimate stress when they are at least three syllables long, and otherwise have penultimate stress. When weak stems are suffixed, the consonant of the weak syllable ([ʈʂ], [k], or [n]) may alternate with another consonant. Patterns of alternation are summarized in Table 3.2, using the active and passive forms of verbs. Note also that weak stems account for a significant portion of the lexicon; a survey of the MDEM shows that around 53% (n = 1302/2424) of active-passive pairs are weak stems.

Here, 'passive' refers to verbs which select a genitive complement, and are built by directly affixing roots (Keenan & Polinsky 2017). The passive suffix has two allomorphs /-ana/ and /-ina/; the distribution of /-ana/ and /-ina/ is said to be lexically specified (i.e. not phonologically predictable), and many verbs have both an /-ina/ and /-ana/ passive.[2] Additionally, many roots have stative passive meaning, and suffixation indicates an action that has not reached its natural cumulation point (e.g. [ˈfaⁿtaʈʂa] 'known' vs. [fanˈtar-ina] 'is (becoming) known' (Keenan, pc).

In addition to these alternants, the lexicon also contains a few minority patterns, such as stems where final [ʈʂa] alternates with [s]. I exclude these because they are so low in frequency that they do not affect my analysis. In the suffixed forms, the final vowel of the

---

[1]The final vowel of weak stems is often devoiced or reduced; this falls out from the general tendency of unstressed vowels to devoice (Howe 2021).

[2]In addition, as will be noted below, some roots surface with an epenthetic 'thematic' consonant under suffixation, e.g. [funu]∼[funusina] 'be enveloped'.

weak stem is not present, leaving the alternating consonant at a morpheme boundary. As demonstrated in these examples, suffixation also shifts stress one syllable to the right.

Malagasy has three suffixes which trigger weak stem alternations: the passive (/-ina, -ana/), the passive imperative (/-i, -u/), and the active imperative (/-a/) (Parker 1883).[3] For example, weak stems alternations are observed before the passive imperative suffix in forms like ['vahuʈʂa]∼[va'hur-i] 'perplexed'. All weak stem alternants included in this study are listed with at least two of these suffixes. In general, weak stem alternations are consistent across the entire morphological paradigm. In other words, weak stem alternations are triggered by suffixation in general and are not restricted to specific suffixes. For simplicity, throughout this chapter, I will use the active and passive forms as representative examples.

pattern		active (m+stem)	passive (stem+ana)	
na ∼	n	man'dʐavina	andʐa'vinana	'to bear leaves'
	m	ma'nandʐana	a'ndʐámana	'to try'
ka ∼	h	ma'ngataka	anga'tahana	'to ask for'
	f	ma'nahaka	ana'hafana	'to scatter'
ʈʂa ∼	r	miánaʈʂa	ianárana	'to learn'
	t	ma'nandʐaʈʂa	ana'ndʐatana	'to promote'
	f	ma'ndʐakuʈʂa	andʐa'kufana	'to cover'

Table 3.2: Patterns of consonant alternation in Malagasy weak stems

3.1.3   A phonological analysis of weak stem alternations

The weak stem alternation pattern can be characterized as consonant neutralization in unsuffixed forms. For example, the two stems in (24) have contrastive pre-suffixal consonants in the passive form ([t] vs. [r]), but both consonants neutralize to [ʈʂ] in the unsuffixed form. Notably, the alternating consonant is prevocalic in both the unsuffixed and suffixed forms; consonant neutralization in this type of prevocalic environment is unusual cross-linguistically, as neutralization typically targets word-final positions (Hayes

---

[3]Allomorphy of the passive imperative (/-i/∼/-u/) is completely predictable, wih /-i/ surfacing as /-i/ after stems containing [u] unless a front vowel intervenes (Zymet 2020).

2011). As discussed in the next section, weak stem alternations make sense if we look at the diachronic changes leading to their development. In this section, I present the standard formal analysis adopted for weak stem alternations.

(24) *Weak stem alternations as neutralization*

STEM	PASSIVE	GLOSS
ˈesuʈ͡ʂa	eˈsur-ina	'remove, take away'
ˈevuʈ͡ʂa	eˈvut-ina	'bounce back, retract'

The standard analysis for weak stems is that they are underlyingly consonant-final (Erwin 1996; Albro 2005). Final consonant neutralization, followed by vowel epenthesis to remove surface codas, are used to derive the weak stem pattern. For example, the stem in [m-iˈanaʈ͡ʂa]∼[iaˈnar-ana] would have the stem UR /ianar/, with surface forms derived as in (25).

(25) Rule-based derivation of ʈ͡ʂa weak stems.

	UR	/m-ianar/	/ianar-an/
Penultimate stress assignment		miˈanar	iaˈnaran
Final C neutralization (/r/→ʈ͡ʂ/_#)		miˈanaʈ͡ʂ	iaˈnaran
Vowel epenthesis (∅→a/C_#)		miˈanaʈ͡ʂa	iaˈnarana
	SR	[miˈanaʈ͡ʂa]	[iaˈnarana]

First, all words are assigned penultimate stress (which is the dominant stress pattern in Malagasy). Following this, the stem-final consonant is neutralized to [ʈ͡ʂ], [k], or [n]. In example (25), /r/ neutralizes to [ʈ͡ʂ] word-finally, but is preserved in the suffixed form, where /r/ is medial and therefore protected from neutralization. Finally, an epenthetic /a/ is added to resolve the violation against codas, counterbleeding final-C neutralization.

Antepenultimate stress falls out naturally from the rule ordering, since penultimate stress assignment precedes vowel epenthesis. As discussed in Albro (2005), weak stems also behave differently from "true" vowel-final stems in compounds; the analysis of weak stems as underlyingly consonant-final can explain these differences. Note that, as will

become clear in the following section, the analysis just given is in fact a recapitulation of the historical development of weak stem alternations.

In modeling Malagasy weak stem reanalysis (§3.3), I adopt this account. Therefore, weak stems are assumed to be underlyingly consonant-final. The URs corresponding to each alternation pattern, taken from Albro (2005), are given in (26). Note that I assign both ʈʂa~f and ka~f forms underlying /f/; the distinction between the two must therefore be made through additional mechanisms such as lexical listing or indexation. Another approach, taken up by prior analyses, is to posit abstract phonemes. For example, Albro (2005) and O'Neill (2015) both analyze ʈʂa~f forms as ending in /ɸ/, which surfaces as [f] in suffixed forms but to [ʈʂ] word-finally.

(26) *Summary of weak stem URs*

pattern	UR	example	
na~n	/n/	[m-anˈɖʐavina]~[anɖʐaˈvin-ana]	/anɖʐavin/
na~m	/m/	[m-anˈanɖʐana]~[anaˈnɖʐam-ana]	/anɖʐavin/
ka~h	/h/	[m-aˈngataka]~[angaˈtah-ana]	/angatah/
ka~f	/f/	[m-aˈnahaka]~[anaˈhaf-ana]	/anahaf/
ʈʂa~r	/t/	[miánaʈʂa]~[ianár-ana]	/ianar/
ʈʂa~f	/t/	[m-aˈnanɖʐaʈʂa]~[anaˈnɖʐat-ana]	/ananɖʐat/
ʈʂa~f	/f/	[m-aˈnɖʐakuʈʂa]~[anɖʐaˈkuf-ana]	/anɖʐakuf/

3.1.4 Historical development of weak stem alternations

In this section, I describe the regular sounds changes which led to the development of weak stems in Malagasy. This description summarizes findings from a large body of scholarship on the history of Malagasy, including Dahl (1951), Hudson (1967), Mahdi (1988), Adelaar (2012), and Adelaar (2013).

Malagasy weak stem alternations started as a series of relatively common final consonant neutralizations, which were subsequently obscured by a process of final vowel epenthesis. Vowel epenthesis was motivated by a phonotactic restriction against codas

which developed around 400AD, when speakers of proto-Malagasy migrated from Kali-mantan into the Comoro Islands. Contact with Bantu during this migration significantly influenced Malagasy grammar, and is thought to have caused the development of final open syllables in Malagasy. For most final consonants, epenthesis of a final vowel re-moved final codas, resulting in the weak stems of modern-day Malagasy.

The development of Malagasy from Proto-Austronesian (PAn) can be broadly be split into three stages: Proto-Malayo-Polynesian (PMP), Proto-Southeast Barito (PSEB), and Proto-Malagasy (PMlg). The examples in (27) and (28) trace a subset of weak stems through these stages, to illustrate the historical development of some weak stem alterna-tions.

(27) *Historical basis of final ʈʂa alternations; changes relevant to the consonant alternation are given in parentheses.*[4]

    a. ʈʂa∼t alternation[5]

PMP	*yawut	*piyawutan	
PSEB	*ˈawut	*piaˈwutan	
PMlg	*ˈavuʈʂ	*fiaˈvutan	(Final affrication, *-t > -ʈʂ)
	*ˈavuʈʂa	*fiaˈvutana	(Final V epenthesis)
Mlg	ˈavuʈʂa	fiaˈvutana	'to uproot'

    b. ʈʂa∼r alternation

PMP	*bukiD	*bukiD-ən	
PSEB	*ˈwukit	*wuˈkiDən	(Final devoicing, *-D > *-t)
	*ˈwukit	*wuˈkirən	(Lenition, *D, *d > r)
PMlg	*ˈwukiʈʂ	*wuˈkirən	(Final affrication, *-t > *-ʈʂ)
	*ˈwukiʈʂa	*wuˈkirəna	(Final V epenthesis)
Mlg	ˈvuhiʈʂa	vuˈhirina	'to make convex'

---

[4]Stress becomes non-contrastive and uniformly penultimate in PSEB; later on, epenthesis of a final vowel resulted in forms with antepenultimate stress, making stress contrastive.

[5]Protoforms use the orthographic conventions established by Dyen (1951). The phonetic value of *R is thought to be [ʀ], *C to be [cç], *y to be [j], *D to be [ɖ].

Example (27a) illustrates the development of a ʈʂa~t alternating weak stems, which historically end in voiceless coronal stops, in this case *t. Final *-t neutralized to *-ʈʂ in PMlg; this affected the non-suffixed forms, while stem-final [t] was preserved in suffixed forms. Following this, epenthesis of a final vowel resulted in the current ʈʂa~t alternation.

In (27b), on the other hand, the PMP stem ends in *D, which is thought to be phonetically [ɖ]. In the non-suffixed form, this final consonant devoiced to *-t, and then neutralized to *ʈʂ. In the suffixed form, *D lenited to [r] due to regular sound change (*D>r); Adelaar 2012). Following this, final vowel epenthesis took place, resulting in the Malagasy ʈʂa~r alternation pattern. Note that while final devoicing (*-D >-t) and lenition (*D > r) are both thought to have taken place in PSEB, devoicing must have preceded lenition for the observed alternations to be possible.

Examples (28a-28b) provide similar illustrative cases for ka-final alternations. First, in PMlg, historical *k spirantized to h intervocalically (before the epenthesis of final vowels). This resulted in ka~h alternations, as shown in (28a). The development of ka~f alternating follows from a similar process, given in (28b). First, *-p and *-k neutralized to -k word-finally. This was followed by spirantization of *p>f.

(28)   *Historical basis of final ka alternations.*

    a. ka~h alternation

PSEB	*ˈtətək	*təˈtək-ən	
PMlg	*ˈtetek	*teˈtehen	(spirantization, *k > h/_V)
	*ˈteteka	*teˈtehena	(Final V epenthesis)
Mlg	ˈtetika	teˈtehina	'to cut into small pieces'

    b. ka~f alternation

PMP	*heyup		
PSEB	*ˈtiup	*pi-tiˈup-an	
PMlg	*ˈtiuk	*pitiˈupan	(Final stop neutralization, *-p >*-k)
	*ˈtiuka	*fitsiˈufana	(Final V epenthesis; spirantization, *p > f/_V)
Mlg	ˈtsiuka	fitsiˈufana	'to lick'

stem-final	alt.	example	PMP/PAn
n	n	ˈankina~aˈnkin-ina	<*n, *ŋ, *l
	m	aˈmpirina~ampiˈrim-ana	<*m
tr	r	ˈampatra~ aˈmpar-ana	< *r,*j [gʲ],*d,*D [ɖ]
	t	ˈharatra~ haˈrat-ana	< *t, *C [cç]
	f	ˈdiditra~ diˈdif-ana	< *p,*b
k	h	baˈliaka~ibaliˈah-ana	<*k,*g
	f	ˈhirika~ hiˈrif-ana	<*p,*b

Table 3.3: Weak stem alternants and corresponding historical consonants

Table 3.3 summarizes all the expected weak stem alternants in Malagasy, given the historical final consonants in PMP. In general, the historical origin of weak stems are well-understood, and the observed alternants in modern Malagasy are expected to correspond to specific historical final consonants.

As a caveat, most consonant-final PMP forms reflect as weak stems in Malagasy, but there are a few exceptions. First, PMP *s, *q, *h were deleted in all environments in PSEB, so do not result in consonant alternations. Additionally, PMP glides *w,*y [j] deleted or coalesced with the preceding vowel in final position, and hardened to *v and *z elsewhere. Stems with a historic final glide therefore reflect as ∅~C alternations in modern Malagasy (e.g. [ˈlalu~laˈlu**v**-ana] < *lalaw, 'pass without stopping'). Additionally, in a few words, final *-n and *-f deleted, which also resulted in ∅~C alternations (e.g. [ˈteni]~[teˈne**n**-ina] 'to be spoken to'). Finally, *s in early Malay loanwords were deleted word-finally, but retained in other positions. These forms have ∅~s alternation in modern Malagasy (e.g [miˈlefa~leˈfa**s**-ana] < *ləpas (Malay) 'gone, escaped'). The reflexes of different PMP final consonants are summarized in Table 3.4.

It should be noted that these ∅~C alternating forms are much less common than weak stems in modern-day Malagasy; they account for around 7% (n=169/2425) of attested stem-suffix pairs in the MDEM.

Coda resolved by	PMP cons.	Mlg alternation	Example
Vowel epenthesis	*-k,*-g	ka~h	baˈliaka~ibaliˈah-ana
	*-p, *-b	ka/ʈʂa~f	ˈhirika~ hiˈrif-ana
	-t,-c	ʈʂa~t	ˈharatra~ haˈrat-ana
	*-d, *-D,*-j	ʈʂa~r	ˈampatra~ aˈmpar-ana
	-n,-ŋ,*-l	na~n	ˈankina~aˈnkin-ina
	*-m	na~m	aˈmpirina~ampiˈrim-ana
Deletion/coalescence	*-y [j]	∅~z	ˈalu~aˈluz-ina
	*-w	∅~v	ˈlalu~laˈluv-ana
Deletion	*-s (loan)	∅~s	miˈlefa~leˈfas-ana

Table 3.4: Malagasy reflexes of stem-final PMP consonants

3.2 Reanalysis in weak stems

Although the historic basis of weak stems is relatively well-understood, there are many mismatches between the observed and expected alternants in Malagasy (given the historic PMP consonant), suggesting that substantial reanalysis has occurred. In the following section, I discuss the predicted outcome of reanalysis under a frequency-matching approach, and show that reanalysis in Malagasy differs from these predictions.

Reanalysis of weak stems in Malagasy always results in the suffixed forms being changed. However, reanalysis may still vary in terms of which alternants are more likely to be reanalyzed, and which alternants are the preferred output of reanalysis.

For example, final [ʈʂa] can alternate with [t], [r], or [f] in the suffixed form. Given these possible alternants, one possible direction of reanalysis is t→r, where a ʈʂa~t alternating stem is reanalyzed as r-alternating. Conversely, reanalysis could happen in the opposite direction, where a historically ʈʂa~r alternating stem becomes t-alternating. (29) summarizes the possible outcomes of reanalysis, given the hypothetical ʈʂa-final weak stem [ˈpakuʈʂa].

(29) *Possible directions of reanalysis for ʈʂa-final weak stems (example stem: [ˈpakuʈʂa])*

Direction	passive (stem + ana)
t → r	pakut-ana→pakur-ana
t → f	pakut-ana→pakuf-ana
r → t	pakur-ana→pakut-ana
r → f	pakur-ana→pakuf-ana
f → t	pakuf-ana→pakut-ana
f → r	pakuf-ana→pakur-ana

Existing work on Malagasy weak stems suggests that in modern Malagasy, the identity of a weak stem's alternant depends not just on the historical consonant, but also various phonological tendencies. Mahdi (1988), in one of the most comprehensive studies of Malagasy weak stems, notes the following generalizations. First, na-final weak stems usually alternates with [n], but may alternate with [m] if the stem-final consonant was historically *m.

Final ka usually alternates with [h], but may alternate with [f] if the historical stem-final consonant was labial, or if the nearest consonant in the stem is [h]. In other words, alternation in ka-final weak stems is partially driven by a dissimilative pattern.

For final [ʈʂa], Mahdi again finds a dissimilative effect. Specifically, the preferred alternant is [r], but that the alternant may be [t] if the stem-final consonant is historically [t], or if there is an [r] somewhere in the preceding stem. Finally, there are also a few words in which -ʈʂa alternates with [f]; these stems all historically end in *p or *b.

Mahdi's findings (and existing work on Malagasy weak stems) have noted the connection between Malagasy alternants and their historical consonant. However, they have not focused on exactly what direction reanalysis happened in, or why there is so much mismatch between the historical consonant and observed alternant in modern-day Malagasy. In this section, I build on Mahdi's work and examine the directions of reanalysis in Malagasy weak stems in detail.

Evidence for reanalysis comes from comparison of historical and modern Malagasy

data. Historical data is taken from the Austronesian Comparative Dictionary (ACD; Blust & Trussel 2010) and Adelaar (2012). Protoforms were excluded if they had less than 6 cognates, and if they didn't reconstruct back to PMP or PAN (i.e. forms were excluded if they were only reconstructable back to Proto-Western Malayo-Polynesian). Additionally, PAN protoforms were only included if they had PMP reflexes.

Modern Malagasy words are taken from the Malagasy Dictionary and Encyclopedia of Madagascar (MDEM; de La Beaujardière 2004), which is an online dictionary that compiles data from multiple Malagasy dictionaries.[6]

§3.2.1 will discuss the distribution of final obstruents in PMP, and what this predicts about the direction of reanalysis in Malagasy. These predictions are compared to the actual observed directions of reanalysis in §3.2.2. §3.2.3 provides additional indirect evidence on how reanalysis has occurred using data from modern Malagasy.

3.2.1 Predicted reanalyses under a frequency-matching approach

In a purely frequency-matching model of morphophonological learning, reanalysis will tend to be in the direction of the more frequent alternant (subject to phonological conditioning). The alternants predicted under this approach can be approximated by looking at the distribution of final consonants in PMP, before extensive reanalysis had taken place.

Table 3.5 shows the distribution of all PMP protoforms with final consonants which would be reflected as weak syllables in Malagasy (n = 805). Results are organized by which alternant each PMP final consonant would correspond to.

There is one complication when [f] is the alternant. Historically, final *-p and *-b neutralized to either *-k or *-t, with a slight bias towards *k (Dahl 1951; Adelaar 2012). Consequently, PMP forms ending in a labial stop tend to reflect as ka-final weak stems, but also often reflect as ʈʂa-final weak stems. In Table 3.5, all PMP forms ending in labial stops

---

Type	alternant	count	P	Predicted reanalysis
ka	h (<*k)	183	0.81	f→h
	f (<*p,*b)	42	0.19	
na	m (<*m)	35	0.10	m→n
	n (<*n,*ŋ)	302	0.90	
ʈʂa	r (<*j,*r,*d,*ɖ)	52	0.25	r→t
	t (<*t)	162	0.75	

Table 3.5: Expected distribution of Malagasy weak stem alternants, based on the distribution of PMP final consonants.

are assumed to correspond to ka-final weak stems in Malagasy. This simplification should not impact the analysis, since ʈʂa∼f alternating forms make up a very small proportion of ʈʂa-final weak stems (n=7,≈2.4%).

From this data, we see that ka-final weak stems have more h-alternating forms, na-final weak stems have more non-alternating forms, and ʈʂa-final weak stems have more t-alternating forms. A frequency-matching approach predicts that reanalysis should generally be in the direction of these more frequent alternants. For example, reanalyses of ʈʂa-final stems should be in the direction of r→t, rather than t→r. Predictions are summarized in the rightmost column of Table 3.5.

Mahdi's (1988) findings on dissimilatory effects in weak stems are also partially replicated in the PMP data. Consider (30), which summarizes the protoforms corresponding to ʈʂa-final stems by whether or not there is a preceding (non-final) [r]. PMP *r ,*d, and *j (in non-final position) are coded as corresponding to Malagasy [r], but excluded if they occurred as the first consonant in a CC cluster. This is because consonant clusters were historically simplified in PMP by deleting the first consonant (e.g. vavaʈʂa, <*bajbaj).

From this data, there appears to be evidence for r-dissimilation. Out of the 28 protoforms coded as containing a preceding [r], only one would reflect as [t]-alternating in Malagasy. In other words, of the 52 forms where the expected alternant is [r], only one was coded as containing a preceding [r].

(30)

	Does stem have [r] (<*r,*d,*ḍ,*j)?	
alternant	yes	no
t	27	136
r	1	51

For ka-final weak stems, however, the evidence for a dissimilatory pattern in PMP is weaker. If dissimilation were present, we would expect the proportion of stems with an immediately preceding *k (corresponding to [h] in modern Malagasy) to be smaller when the expected alternant is [h]. When the expected alternant is [h], around 8% (n = 13/147) of protoforms have a preceding *k. When the expected alternant is [f], 22% of forms (n = 13/60,22%) have a preceding *k. In other words, there is a slight dissimilatory pattern, but it is weaker than the r-dissimilation pattern observed in (30).

(31)

	does stem have h (<*k)?	
alternant	yes	no
h	13	134
f	13	47

3.2.2 Observed directions of reanalysis

In this section, I discusses form-by-form comparisons of PMP stems to their weak stem reflexes. Where there is a mismatch between PMP and Malagasy, the direction of reanalyses can be inferred. The ACD contains 143 protoforms that reflect as productive suffixed forms in Malagasy. 56 were removed following the exclusionary criteria discussed above, leaving 87 forms to be analyzed. The data is also supplemented with 49 Malay and Javanese loanwords from the World Loanword Database (WOLD; Adelaar 2009) and Adelaar (1994). These are all early loans, introduced to Malagasy before the development of weak stems (Adelaar 1989). Tables 3.6-3.8 summarize whether the alternant observed in Malagasy matches the expected one given the historical consonant (or in the case of loanwords, the final consonant of the source word).

Table 3.6 shows the results for na-final weak stems. The column named 'PMP' shows the expected alternant given the PMP protoform, while the column named 'Mlg' shows the actually observed alternant in Malagasy. Mismatches between PMP and Mlg indicate that a reanalysis has occurred. Overall, there are relatively few reanalyses (n=3), but most are in the direction of m→n (e.g. [ˈlalina~lalˈin-ina] <*dalem 'inside, deep'). This is in line with the predictions of an inductive approach.

PMP	Mlg	Match?	Count
m	m	yes	2
	n	no (m→n)	2
n	n	yes	38
	m	no (n→m)	1

Table 3.6: Expected (PMP) vs. observed (Malagasy) alternant of na-final stems, based on known protoforms/loanwords

Of the stems expected to be n-alternating, only one has been reanalyzed in the direction of n→m (1/39, 3%); the reanalyzed stem is [ˈtenona~teˈnom-ina] (<*tenun) 'to weave/be woven'. Given the lack of data, it is hard to tell what the cause is.[7] Overall, comparisons for the na-final weak stems are tentatively in line with a statistical learning approach.

Table 3.7 shows the reanalyses for ka-final weak stems. Once again, there are relatively few cases of reanalyses (n=2). However, both case of reanalysis are in the direction of f→h (e.g. [ˈaʈʂika~fiaˈʈʂeh-ana] <*qadep 'face, facade'), in line with the predictions of a frequency-matching approach. In contrast, there are no reanalyses in the direction of f→h.

Note that the data did not contain any stems where the immediately preceding consonant is [h]. As such, it is unclear whether a dissimilatory effect was active in the reanalysis of ka-final weak stems. However, one item, which was excluded because it

---

[7]This change of n→m does not seem to be from a dissimilatory effect, since there was no nasal dissimilation found in either PMP or modern Malagasy. However, nasal dissimilation is documented the Betsimisaraka dialect of Malagasy (O'Neill 2015)

PMP	Mlg	Match?	Count
f	f	yes	3
	h	no (f→h)	2
h	h	yes	36
	f	no (h→f)	0

Table 3.7: Expected vs. observed alternant of ka-final stems, based on known proto-forms/loanwords

was only reconstructed to PWMP (Proto-Western Malayo-Polynesian), shows reanalysis in the direction of h→f that could potentially be attributed to h-dissimilation. This word, [ˈlauka~laˈufana] (<PWMP *lahuk) 'meat/relish eaten with rice', historically had a preceding [h] which was subsequently elided in PSEB.

Table 3.8 shows results for ʈʂa-final weak stems. The rightmost column, 'has r?', indicates, for each row, the number of forms which had an [r] in the stem. For ʈʂa-final stems, extensive reanalysis has occurred towards [r]. Of the stems that were historically expected to have [t] as the alternant, over half (23/40, 57%) have been reanalyzed in the direction of t→r (e.g. [ˈhudiʈʂa~huˈdir-ina] <*kulit, 'skin, hide'). In contrast, when the expected alternant is [r], there is only one case of reanalysis (n=1). Moreover, the one case of reanalysis in the ʈʂa~f alternating forms is in the direction of f→r ([ˈhalaʈʂa~aŋaˈlar-ina] <*alap, 'theft, robbery').

PMP	Mlg	Match?	Count	has r?
t	t	yes	17	7 (41%)
	r	no (t→r)	23	0
	f	no (r→f)	0	0
r	r	yes	11	0
	t	no (r→t)	1	1
	f	no (f→t)	0	0
f	f	yes	3	1 (33%)
	t	no (f→t)	0	0
	r	no (f→r)	1	0

Table 3.8: Expected vs. observed alternant of ʈʂa-final stems, based on known proto-forms/loanwords

Additionally, r-dissimilation appears to be active in the reanalysis of ʈʂa-final weak

stems, in that reanalysis to [r] is blocked if the stem has a preceding [r]. As seen in Table 3.8, when the alternant was reanalyzed to be [r], the stem never contained a preceding [r]. In addition, out of the t-alternating stems that were not reanalyzed, a relatively larger proportion (n=7/17, 41%) had a preceding [r] (e.g.[ˈuɻiʈʂa∼uˈɻitana] <*qurit, 'stroke, line').

The only example of reanalysis in the direction of r→t is likely also motivated by r-dissimilation. The reanalyzed form [ˈsanɖʐaʈʂa∼anaˈnɖʐat-ana] (<sandar, Malay loan) does not have a preceding [r] in modern Malagasy, but [nɖʐ] sequences are historically [nr], and only affricated to [nɖʐ] in a later stage of PSEB (Proto Southeast-Barito).

The direction of reanalysis in ʈʂa-final weak stems goes against predictions of an inductive approach. Based on the PMP distribution, there should more [t]-alternating forms than [r]-alternating forms. However, reanalyses are overwhelmingly towards the less frequent alternant, in the direction of t→r.

3.2.3 The result of reanalysis: weak stem alternations in modern Malagasy

This section describes the distribution of weak stem alternants in modern Malagasy, using 1893 stems taken from the MDEM. This data supplements the above results, by providing indirect evidence for the direction of reanalysis that has taken place.

ending	alternant	Freq	
ka	h	668	(94.8%)
	f	35	(5.0%)
	other	2	(0.2%)
na	n	580	(97.7%)
	m	13	(2.2%)
	other	1	(0.1%)
ʈʂa	r	231	(70.2%)
	t	89	(27.1%)
	f	7	(2.1%)
	s	2	(0.6%)

Table 3.9: Proportion of alternants for modern Malagasy weak stems

Table 3.9 summarizes the distribution of weak stem alternants in modern Malagasy. The na-final weak stems are overwhelmingly non-alternating, where 97.7% of the sampled forms are non-alternating. This distribution suggests that reanalyses have been in the direction of m→n, increasing the relative frequency of non-alternating na-final weak stems. Note that one na-final weak stem, ['biana]∼[bi'anina, bi'afina], has two listed alternants [n] and [f]. This form is labeled 'other' in Table 3.9.
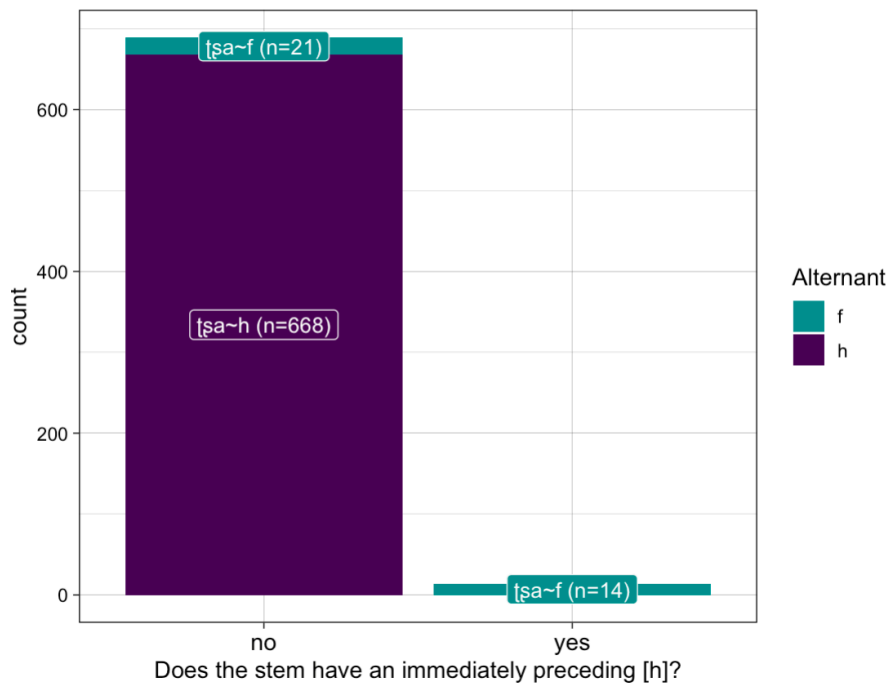
Figure 3.1: Distribution of alternants in ka-final weak stems

For ka-final weak stems, [h] is overwhelmingly the preferred alternant, accounting for 94.8% of the sampled forms. Again, this distribution is consistent with the finding that reanalyses have been in the direction of f→h.

In addition, recall that Mahdi (1988) finds evidence for h-dissimilation in ka-final weak stems. A weak potential effect of h-dissimilation was found in PMP. Consistent with this, h-dissimilation does seem to be present in modern Malagasy. This is illustrated in Fig. 3.1, which shows the distribution of alternants for ka-final stems by whether the consonant nearest to the alternant is [h]. When there is an immediately preceding
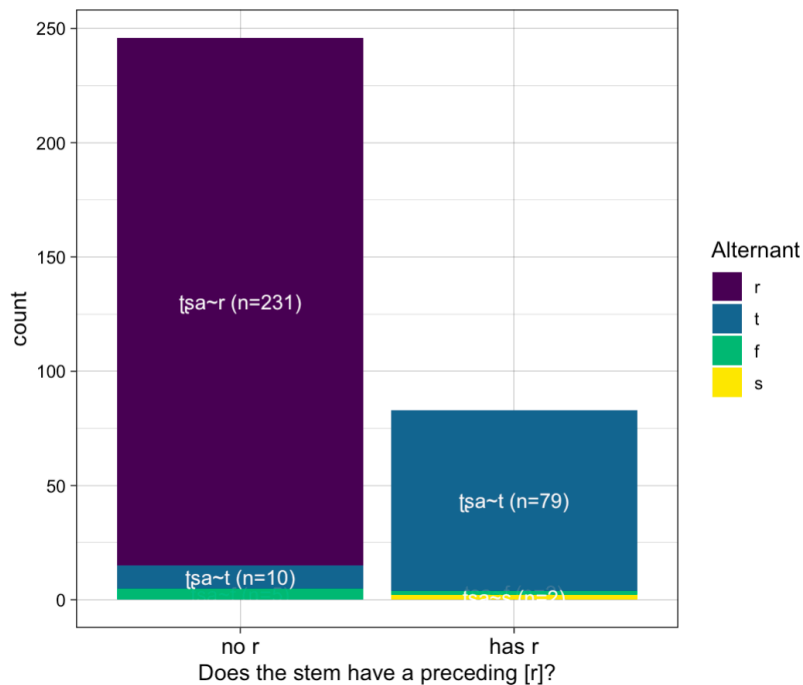
Figure 3.2: Distribution of alternants in ʈʂa-final weak stems

[h], the observed alternant is always [f]. In contrast, when the stem does not have a preceding [h], only 3% (n = 21/689) stems have [f] as the alternant. Based on these results, h-dissimilation could have affected reanalyses of ka-final stems. However, results are tentative because there are very few f-alternating stems, and also very few stems which contain a preceding [h].

The data in Table 3.9 shows that for ʈʂa-final stems, there is a general preference for alternation with [r] (relative to [t] or [f]), such that around 70.2% (231/329) of relevant stems are r-alternating. Fig. 3.2 shows the proportion of alternants, organized by whether or not there is a preceding [r] somewhere in the stem. From here, it is evident that in modern Malagasy, there is a strong f-dissimilation pattern.

Specifically, final ʈʂa *never* alternates with [r] if there is already an [r] in the stem. In contrast, when the stem has no preceding [r], there is a strong, near-exceptionless preference for alternation with [r]. As seen in the 'no r' condition in Fig. 3.2, once dissimilatory effects are accounted for, ʈʂa-final weak stems alternate with [r] around

94% of the time (n = 231/247). Overall, the distribution of alternants in modern Malagasy supports the finding that reanalysis in ʈʂa-final weak stems has been in the direction of t→r, except when blocked by r-dissimilation.

3.2.4 Markedness effects on the reanalysis of ʈʂa stems

For the ʈʂa-final weak stems, reanalysis in the direction of t→r cannot be explained by a frequency-matching approach. Additional factors are needed to explain this direction of reanalysis.

I propose that reanalysis towards [r] is the result of a markedness bias in Malagasy against intervocalic stops. There is support for the presence of this constraint internal to the Malagasy lexicon. Historically, Malagasy underwent intervocalic lenition which affected all stops except for *t (*b>v, *p>f, *d,*ɖ>r, *k,*g>h) (Adelaar 1989, 2012). As such, it's likely that there were very few intervocalic stops at some point in an earlier stage of Malagasy.

A constraint against intervocalic stops is also independently motivated cross-linguistically. Studies have found phonetic support for intervocalic lenition, from both an articulatory (Kirchner 1998) and perceptual (Kaplan 2010; Katz 2016) point of view. There is also sizeable typological support for intervocalic lenition being extended across morpheme boundaries, including (among many other examples) Sanskrit stop voicing (Selkirk 1980), English phrasal tapping (Hayes 2011, p. 143-144), Korean lenis stop voicing (Jun 1994), and Catalan fricative weakening (Wheeler 2005, p. 163). Malagasy ʈʂa~r alternation fits into this typology, and can be characterized as stop lenition at morpheme boundaries.

The fact that only ʈʂa-final stems, and not other weak stems, have undergone reanalysis in a direction not predicted by distributional information, follows naturally from this markedness-based account. For ka-final stems, the possible alternants are [f] and [h]; both are fricatives and would not violate a constraint against medial stops. For na-final stems, the attested alternants are [m] and [n]. Both violate a constraint against medial

stops, so are equally marked if all else is held equal. Existing surveys of lenition also find that intervocalic nasals are more stable than their obstruent counterparts, and therefore presumably less marked (Kirchner 1998; Lavoie 2001).

In Chapter 1, a distinction was made between 'active' and 'universal' markedness, where active markedness effects are already present in the phonology of the language, namely as stem phonotactics. I proposed that the markedness effects present in reanalysis are restricted to such active effects. As will be discussed in §3.3.5, it turns out that the Malagasy data is consistent with this proposal, as constraints penalizing intervocalic stops are active in the stem phonotactics.

Finally, it is worth noting that the pattern of r-dissimilation, though already present in the distributional information, also has typological support. Suzuki (1998), in a typological study of dissimilation, finds multiple examples of tap dissimilation. More generally, liquid dissimilation is also crosslinguistically attested, both as a phonotactic tendency and in active phonological processes (e.g. French and Spanish; Colantoni & Steele 2005).

3.2.5 Other conditioning factors

Existing work on paradigm learning, discussed in Chapter 1, shows that learners are sensitive to statistical sub-generalizations within paradigms; this idea is built into some rule-based models (e.g. Albright & Hayes 2003), and is also implied in analogical models (e.g. Ernestus & Baayen 2003; Nosofsky 2011). For example, English speakers learn a general rule of English past tense formation, but also learn sub-generalizations for words such as *dive/dove* and *strive/strove*.

One possibility, not considered so far, is that reanalysis in Malagasy is actually the result of speakers extending certain subgeneralizations to the data. If this is the case, then in modern Malagasy, the distribution of t- and r-alternating forms should be conditioned by phonological factors within the weak stem paradigm. Here, I explore two possible factors—identity of the preceding vowel and the passive allomorph that a stem selects—

71

and show that neither are strong predictors of tra-final weak stem alternants in modern Malagasy.

First, Table 3.10 shows the distribution of alternants ([t] vs. [r]) for the ʈʂa-final weak stems, based on the identity of the vowel immediately preceding the weak syllable. We can see that across each vowel category, the distribution of alternants is relatively even, with /t/ occurring around 73-85% of the time.

Final V	alternant	n	p	Example
a	r	61	0.73	[ˈsulaʈʂa]~[suˈlarana]
	t	22	0.27	[ˈfaraʈʂa]~[faˈratana]
e	r	29	0.85	[ˈfeʈʂa]~[feˈrana]
	t	5	0.15	[paˈrareʈʂa]~[paraˈretina]
i	r	58	0.73	[ˈvusiʈʂa]~[vuˈsirana]
	t	21	0.27	[ˈuritra]~[uˈritana]
u	r	66	0.77	[ˈhusuʈʂa]~[huˈsurana]
	t	20	0.23	[uruʈʂa]~[uˈrutana]

Table 3.10: Distribution of ʈʂa weak stem alternants by vowel

Recall also that passive suffix has two allomorphs /-ana/ and /-ina/, whose distribution is described as being unpredictable. I also compare the distribution of alternants ([t] vs. [r])against the choice of allomorph (/-ana/ vs. /-ina/), and find no clear pattern. The results, shown in Table 3.11, show that for both suffix allomorphs, /r/ is the observed alternant around 74-79% of the time.

Suffix	Alt	n	p
ana	r	79	0.74
	t	28	0.26
ina	r	139	0.79
	t	37	0.21

Table 3.11: Distribution of ʈʂa weak stem alternants by passive allomorph

3.2.6 Alternative accounts

In this section, I describe two alternative explanations for why reanalysis of ʈʂa-final stems has been in the direction of t→r. Both will also be considered in Section 3.3, where a quantitative model of reanalysis is implemented.

PERCEPTUAL SIMILARITY. One alternative explanation is that speakers are driven by a perceptual similarity bias, rather than a markedness bias (Steriade 2009; Wilson 2006; White 2013). This is the idea that learners prefer perceptually less salient alternations. That is, if [ʈʂ] has a smaller perceptual distance to [r] than to [t], reanalysis towards [r] could be explained as the result of a bias towards perceptually similar alternations.

Although there have been no studies on perceptual distance of Malagasy phonemes, there is indirect evidence from English that [ʈʂ] is perceptually closer to [t] than to [r]. If this is true, than a perceptual distance account predicts that [ʈʂ]∼[t] alternation is preferred over [ʈʂ]∼[r] alternation. English does not phonemically have [ʈʂ] and [r], but Warner et al. (2014) have found that for English, [tʃ] is perceptually closer to [t] than to [ɾ]. If we use [tʃ] and [ɾ] respectively as proxies for Malagasy [ʈʂ] and [r], this would suggest that [ʈʂ] is perceptually more similar to [t] than to [r]. This assumption is not unreasonable because Malagasy [ʈʂ] is variably realized as postalveolar, and [r] is realized as a tap in fast speech (Howe 2021).[8]

TOKEN FREQUENCY. There is a large body of evidence showing that type frequency, rather than token frequency, is a much better predictor of frequency-matching behavior in morphophonological learning (Bybee 1995, 2001; Pierrehumbert 2001; Albright & Hayes 2003, etc.). The model I adopt (and existing probabilistic models like Albright's MGL) are trained on types rather than tokens. However, token frequency can still play an important indirect role in learning. In particular, token frequency can determine what

---

[8]There is also evidence of low discriminability between retroflex and coronal affricates ([ʈʂ] vs. [ts]; [ʈʂʰ] vs. [tsʰ]) in Mandarin Chinese, where the two places of articulation are phonemically contrastive (Cheung 2000; Tsao et al. 2009).

data a learner is likely to encounter; forms with higher token frequency are more likely to be encountered, and can potentially have a larger effect on the grammar.

In the case of Malagasy, we could imagine a scenario where tra~r alternating forms have much higher token frequency than tra~t forms. As a result, they would be more likely to be learned by speakers, and therefore be represented more in the type frequency. In other words, reanalysis of t→r could potentially be explained by statistical learning, if tra~r forms are generally higher in token frequency. In the following section, I explore this possibility, and show that it does predict t→r reanalysis, but at a magnitude that is too small to match the Malagasy data.

3.2.7   Interim summary

Comparison of PMP protoforms with Malagasy suggests that reanalysis of weak stems is driven not just by distributional probabilities of the lexicon, but also by additional markedness effects. Findings of this section are summarized in Table 3.12. On one hand, reanalysis of na- and ka-final weak stems is largely predictable from distributional probabilities.

For the ka-final stems, there is also tentative support for h-dissimilation both in the PMP distribution and in modern-day Malagasy. However, the lack of evidence makes this pattern harder to confirm. For this reason, I do not consider the effects of h-dissimilation in the rest of this chapter.

Type	Pattern	Frequency-matching?
na	m→n	yes
ka	f→h	yes
	h-dissimilation	yes?
ʈʂa	t→r	**no**
	r-dissimilation	yes

Table 3.12: Summary: directions of reanalysis in Malagas

For the ʈʂa-final stems, there was distributional evidence in PMP for (i) t-alternation, and (ii) r-dissimilation. Although the r-dissimilation pattern holds true in modern Malagasy, reanalysis has generally occurred in the direction of t→r, which is the opposite of what is predicted by lexical statistics. In other words, a purely frequency-matching model of reanalysis would fail to predict the direction of reanalysis found in Malagasy.

Instead, reanalysis of ʈʂa-final stems is argued to be driven by a markedness constraint against intervocalic stops. In the following section, I outline a model of reanalysis that incorporates a markedness bias, and show that it better captures the Malagasy data than an unbiased frequency-matching model.

The rest of the chapter will focus on markedness effects in ʈʂa-final weak stems, where the effects of markedness are most pronounced.

3.3   Modeling reanalysis with a markedness bias

In this section, I test the predictions of the previous section (that reanalysis in Malagasy is driven by both distributional and markedness effects) using a quantitative model of reanalysis. As a preview, results in this section explicitly demonstrate that both distributional and markedness effects are needed to explain the direction of reanalysis found in Malagasy.

3.3.1   Components of a model of reanalysis

The model architecture is detailed in Chapter 2, and briefly summarized in this section. The model has three main components. First, it uses MaxEnt (Goldwater & Johnson 2003; Smolensky 1986), a probabilistic variant of Optimality Theory. Additionally, to mirror the effect of reanalyses over time, the model has an iterative (generational) component, in which the output of one iteration of the model becomes the input for the next (see §2.3 for details). Finally, to incorporate markedness effects, a bias is implemented as a

Gaussian prior, following the methodology of Wilson (2006) and White (2013, 2017). This biased model will be compared to control models that do not have a markedness bias.

For explanatory ease, tableaux used to demonstrate the effect of different constraints will be shown in classic strictly ranked OT. However, for the actual model, constraints are weighted and the model output is a set of candidates, each with a predicted probability.

3.3.2 Inputs

In Malagasy, reanalysis happens when speakers do not know the suffixed form of a stem. In other words, they must infer what a stem is underlyingly while faced with structural ambiguity. Reanalysis can therefore be modeled as UR inference, which I describe in detail in §2.2.

Under this approach, the candidate set is UR-SR pairs. For example, suppose that the learner hears a new word ['vukiʈʂa], and want to produce the passive form. The candidate set would include items like "/vukit-an/ [vuˈkitana]" and "/vukir-an/ [vuˈkirana]". UR inference constraints enforce specific UR-SR mappings. For example, a constraint like [ʈʂa#] = /t/ penalizes a final [ʈʂa] which is not underlyingly /t/.

The model was trained on 1270 nonce weak stem-suffixed pairs, designed to represent historical Malagasy, presumably before extensive reanalysis had occurred. Relative frequencies of ka, ʈʂa, and na stems match that of the MDEM corpus. The relative frequency of each alternant was based on the distribution of final consonants in the historical PMP data. Nonce stems are used in place of actual PMP stems because the number of available PMP forms is too few.

The candidate set is constrained to only include UR-SR mappings that are observed in the lexicon. For example, given the input ['vukiʈʂa] + [ana], a candidate like /vuki**p**-an/ [vuˈkipana] is excluded because ʈʂa~p alternation is never observed in the lexicon. In addition, ʈʂa~f alternating forms and irregular alternants (e.g. na~f alternating forms)

are excluded because they are very low-frequency; when included, they do not influence model outcomes. The input data is summarized in Table 3.13.

Input	Candidate	Freq	P
[ˈvukiʈʂa] + [ana]	/vukiʈʂ-an/∼[vuˈkiʈʂana]	0	0
	/vukir-an/∼[vuˈkirana]	56	0.30
	/vukit-an/∼[vuˈkitana]	131	0.70
[ˈvuriʈʂa] + [ana]	/vuriʈʂ-an/∼[vuriʈʂan]	0	0
	/vurir-an/∼[vuˈrirana]	0	0
	/vurit-an/∼[vuˈritana]	56	1
[ˈvukika] + [ana]	/vukik-an/∼[vuˈkikana]	0	0
	/vukih-an/∼[vuˈkihana]	490	0.90
	/vukif-an/∼[vuˈkifana]	57	0.10
[ˈvukina] + [ana]	/vukin-an/∼[vuˈkinana]	440	0.92
	vukim-an/ [vuˈkimana]	40	0.08

Table 3.13: Sample inputs to the Malagasy model of reanalysis

In this implementation, reanalysis is always from a non-suffixed allomorph. An unsuffixed form must serve as the base of reanalysis, since if speakers knew the suffixed form (where the weak stem alternant surfaces), there would be no ambiguity that could result in mis-learning of the weak stem alternation pattern. As noted in Chapter 1, a similarly restricted approach is taken by Albright (2002a; 2010, etc.), who argues that the base of reanalysis is fixed, and is always a single slot of a morphological paradigm.

Albright also argues that the base should be the most **informative** allomorph, which has the most contrastive information. The Malagasy base appears to contradict this hypothesis, since it is the suffixed forms that are more informative, and retain contrastive information about weak stem consonant alternations. The Malagasy data may lead us to slightly rethink Albright's hypothesis that informativeness always determines the base of reanalysis. In particular, the base of reanalysis is generally the most informative one (per Albright's hypothesis). However, if learners only have access to limited paradigm slots, reanalyses may still occur from these paradigm slots even if they are not the most informative.

Token frequency may also affect how learners select the base of reanalysis to some

degree. Albright (2008) suggests that when one slot of the paradigm is used with much higher frequency than others, it may be preferred as the base of reanalysis. However, Keenan & Manorohanta (2001) find, based on written corpora, that actives (unsuffixed) and passives (mostly suffixed) occur at roughly equal rates, making this explanation less likely. Another possible factor is the tendency for bases to be isolation stems or other shorter, 'unmarked' forms (Vennemann 1972; Kuryłowicz 1945).

3.3.3 UR inference and faithfulness constraints

UR inference constraints guide the mapping of SRs to URs. This is demonstrated in (32), which derives the suffixed form of a ʈʂa-final weak stem. Candidate (a), where surface [ʈʂa#] is mapped to /t/, violates the constraints '[ʈʂa#] = /r/' and '[ʈʂa#] = /ʈʂ/'. In contrast, candidate (b) violates '[ʈʂa#] = /t/' and '[ʈʂa#] = /ʈʂ/'. Finally, the candidate (c), which is non-alternating, violates the two constraints which enforce alternation, '[ʈʂa#] = /r/' and '[ʈʂa#] = /t/'.

(32)   *UR inference constraints for ʈʂa-final weak stems*

['vuliʈʂa] + [ana]	[ʈʂa#] = /t/	[ʈʂa#] = /r/	[ʈʂa#] = /ʈʂ/
a. /vulit-an/ vulitana		*	*
b. /vulir-an/ vulir-ana	*		*
c. /vuliʈʂ-an/ vuliʈʂ-ana	*	*	

Note that in this model, alternation is enforced by UR inference constraints. For example, in the above tableau, the constraint [ʈʂa#] = /r/ motivates the learner to posit the UR /vulir/, which corresponds to a ʈʂa~r alternating weak stem.

Derived surface forms can also be in a correspondence relationship with the stem SR (Benua 1995) (i.e. output-output correspondence). To encode these output-output relations, I use the *MAP family of faithfulness constraints instead of classical feature-based faithfulness constraints (McCarthy & Prince 1995). *MAP constraints, proposed by Zuraw (2010b, 2013), assess violations between pairs of surface forms. A constraint *MAP(a, b)

assesses a violation to a candidate if *a* is mapped to a corresponding *b*. The correspond-ing segments *a* and *b* can differ more than one feature. For example, a constraint like *MAP(k,f), where segments [k] and [f] differ in multiple features ([continuant], [LABIAL], [DORSAL]), is allowed.[9]

The tableau in (33) demonstrates how *MAP violations are assessed for a ʈʂa-final weak stem; this tableau is the same as (32) above, with the addition of two *MAP constraints. Candidate (a), where [ʈʂ] alternates with [t], incurs a violation of *MAP(ʈʂ, t). Candidate (b), where the alternant is [r], incurs a violation of *MAP(ʈʂ,r). The faithful candidate (c), on the other hand, incurs no output-output faithfulness violations.

(33)   *MAP *constraints for ʈʂa-final weak stems*

[ˈvuliʈʂa] + [ana]	[ʈʂa#] = /t/	[ʈʂa#] = /r/	*MAP(ʈʂ,t)	*MAP(ʈʂ,r)
a. /vulit-an/ vulitana		*	*	
b. /vulir-an/ vulir-ana	*			*
c. /vuliʈʂ-an/ vuliʈʂ-ana	*	*		

*MAP constraints are more powerful than traditional faithfulness constraints, but are also constrained in substantive terms. Specifically, Zuraw assigns *MAP constraints a default weighting (or ranking) based on the **p-map**. The p-map, proposed by Steriade (2009), is a language-specific perceptual map which encodes the perceptual distance be-tween all segment pairs in all contexts. *MAP constraints which ban changes that cover a larger perceptual distance are assigned a default ranking higher (or weighted more) than constraints banning smaller changes.

In an inductive model of Malagasy, traditional output-output identity constraints actu-ally do just as well as *MAP constraints in frequency-matching the input data. However, the current study adopts *MAP constraints because they more straightforwardly allow dif-ferent types of learning bias to be incorporated. In particular, they have been successful

---

[9]Zuraw also permits *MAP constraints to include contexts. For the present paper, context-free *MAP constraints suffice.

at modeling phonetic similarity bias in prior work (Wilson 2006; Hayes & White 2015). This is important because although my focus is on markedness bias, I also consider an alternative model which incorporates a phonetic similarity bias in §3.3.7.

Alternation of weak stems falls out from the relative weighting of UR Inference constraints and output-output faithfulness constraints. In (33) above, if the UR inference constraints outweigh (or outrank) *MAP constraints, the [ʈʂ] of a weak stem will always alternate with [t] or [r]. On the other hand, if *MAP constraints have higher weights, the grammar will prefer the non-alternating candidate (c).

3.3.4 Markedness constraints

In a typical OT model, alternation is enforced by ranking markedness above faithfulness. However, as described above, in the current model, alternation is enforced by UR inference constraints.

The frequency-matching model of Malagasy reanalysis needs just one additional markedness constraint. This constraint, *r...r], is used to enforce dissimilation of [r] at the right edge of morpheme boundaries.

Reference to morpheme boundaries is necessary because within stems, r...r sequences are allowed (e.g. [ˈraraka] 'spilled', [buˈrera] 'weak, limp', [ˈrirana] 'edge'). This approach is similar to the one taken by Pater (2007) and Chong (2019) to explain morphologically-derived environment effects (MDEEs), where static phonotactic patterns mismatch the alternations allowed at morphological boundaries. The effect of *r...r] is demonstrated in tableau §3.3.4, where the input stem has a preceding [r]. In this tableau, highly ranked *r...r] rules out the r-alternating candidate (b).

(34) *Effect of *r...r]*

ˈvuriʈʂa	*r...r]	*MAP(ʈʂ,t)	*MAP(ʈʂ,r)
☞ a. vurit-ana		*	
b. vurir-ana	*!		*

The model laid out so far is able to match the input data perfectly ($R^2 = 1$). However, the goal of the model is not to fit the input data. Instead, given input data that represents Malagasy before reanalysis, it should predict predict the correct direction of reanalysis, and match the distribution of alternants in modern Malagasy. The current inductive model will not be able to do this, as it predicts reanalysis to be in the direction of high frequency alternants (r→t, f→h, m→n). This makes the wrong prediction for ʈʂa-final stems, where reanalysis is in the direction of t→r.

3.3.5   Learning additional markedness constraints

The central argument of this chapter is that reanalysis in Malagasy is partially driven by markedness effects that *cannot* be learned from frequency distributions within a paradigm. Instead, speakers are also sensitive to more general lexicon-wide markedness effects. In this section and the subsequent section, I outline a process for incorporating this markedness component to the model.

To test whether the relevant constraint is present in Malagasy phonotactics, I constructed a phonotactic model of Malagasy stems using the UCLA Phonotactic Learner (Hayes & Wilson 2008), which learns a grammar of n-gram constraints that fits the distribution of natural classes in a set of learning data. The grammar was restricted to learning maximally trigram-length constraints. The UCLA Phonotactic Learner also allows the user to specify different projections that encode long-distance dependencies. The Malagasy phonotactic grammar included two projections, a vowel tier ([+syllabic]) and consonant tier ([-syllabic]).

The input to the grammar was 3800 Malagasy stems selected randomly from the MDEM. Completely reduplicated forms were automatically removed (e.g. [pakapaka]), but partially reduplicated forms still remain. Only non-suffixed stems were used; suffixed allomorphs were not included because the alternants of weak stems reflect the distribution of the lexicon *after* reanalysis, while the phonotactic grammar is supposed to approximate

patterns already present in Malagasy pre-reanalysis.

The resulting grammar included four constraints, given in (35), that penalize intervocalic stops and specifically favor [r] over [t] as the alternant for ʈʂa-final weak stems. Note that because Malagasy has no codas (and by extension final consonants), a constraint *[+syll]C (where C is any consonant) is equivalent to *[+syll]C[+syll]. The constraints listed here all motivate reanalysis of t→r. Crucially, they also do not affect the relative preference for different alternants in ka- or na-final weak stems.

(35) *Phonotactic constraints penalizing intervocalic stops*

Constraint	Violations
*[+syll][-cont,-vc][+syll]	V{p,t,ts,dz,ʈʂ,k}V
*[+syll][-son,-cont][+syll]	V{p,b,t,d,ts,dz,ʈʂ,ɖʐ,k,g}V
*[+syll][-tap,-nasal,+coronal]	V{t,d,ts,dz,ʈʂ,ɖʐ,s,z,l}V
*[+syll][-son,-cont,-labial]	V{t,d,ts,dz,ʈʂ,ɖʐ,k,g}V

For simplicity, I added only the constraint *V[-cont,-voice]V to the model. Although the phonotactic grammar found multiple constraints which penalize intervocalic stops, I included only one because all four constraints have the same violation profile with respect to the candidates in weak stem reanalysis. *V[-cont,-voice]V is assigned a bias towards higher weight following the method described in Chapter 2. Specifically, σ^2 is set to 1 and μ is varied to implement different learning biases. For example, a markedness bias is implemented by assigning *V[-cont,-voice]V a higher μ than competing faithfulness constraint(s). As a result, *V[-cont,-voice]V will be biased to have a higher weight than the relevant faithfulness constraints. In §3.3.7, I provide the specific μ values used for the markedness-biased model, as well as the μ values of baseline control models.

Note that alternation in ʈʂa-final weak stems is also driven by a strong r-dissimilation constraint. The phonotactic grammar did not learn this constraint in the consonant tier; other projections that were tested, such a CORONAL tier, also did not learn a constraint for r-dissimilation. Constraints on dissimilation of larger classes of segments (e.g. approximants) were also found to be non-significant. As such, r-dissimilation differs from

lenition in that it is a markedness constraint learned from the local distribution of weak stem alternants, and does not receive additional phonotactic support.

In other words, although *r...r] and *V[-cont,-voice]V look similar on the surface, they have different underlying mechanisms. Reanalysis driven by *V[-cont,-voice]V is motivated by stem phonotactics. In contrast, reanalysis driven by *r...r] is better characterized as frequency-matching of patterns within the weak stem paradigm. As discussed in §1.4, markedness-driven reanalysis only arises when there is uncertainty in an alternation pattern. As shown in §3.2.1, even in the PMP input, r-dissimilation is already a near-exceptionless pattern (with only one exception). Because there is strong distributional evidence for r-dssimilation within the paradigm, speakers obey this generalization even in the absence of phonotactic support.

3.3.6 Iterated learning parameters

As described in §2.3, the iterated learning component of the model simulates reanalysis over time by having the output of one generation become the input for the next. At each generation, some proportion of suffixed forms are randomly sampled and forgotten. The Agent learns a grammar based on the remembered forms, and use this to generate new suffixed forms (replacing the forgotten forms). Two parameters can be set in the iterated learning component of the model: forgetting rate (i.e. the proportion of suffixed forms forgotten in each generation) and number of generations.

As discussed in §2.3, increasing the forgetting rate will in general increase the magnitude of markedness effects in each iteration. I tested 5 forgetting rates (0.05, 0.1, 0.15, 0.20, 0.25) and ran the model for 50 generations. In the interest of clarity, and because the model trended in the same direction across all five forgetting rates, the rest of this chapter will only present models with a forgetting rate of 0.2.

The number of iterations was chosen to reflect the maximal span of time in which reanalyses of weak stems could have occurred. The sound changes that resulted in weak

stem alternations took place around 6-7th century AD, while the modern Malagasy data starts around the 1800s. Therefore, reanalysis must have occurred within the span of around 1200 years. This corresponds to roughly 50 generations, assuming that each generation is 25 years. 50 generations is meant as a conservative estimate, since in reality, reanalysis of the ʈʂa-final weak stems may have happened in a much shorter span of time. Note that if reanalysis did in fact occur over a shorter time span, the model can be modified to reflect this by varying parameters such as forgetting rate (see §2.3 for a full discussion of the relevant parameters).

Because random sampling causes each iteration of the model to vary slightly, all subsequent models were run 20 times, and predicted probability values are the mean of these 20 trials.

3.3.7 Model comparison

This section compares markedness biased models against controls (reflecting the alternative analyses discussed above) to evaluate the effect of markedness in improving model predictions.

Four models are compared: BASELINE (flat/uniform prior), P-MAP, TOKEN FREQUENCY, and MARKEDNESS. The priors assigned to each condition are explained below and summarized in Table 3.14 below. All conditions except for TOKEN FREQUENCY assume that inputs have uniform token frequency, and are therefore equally likely to be sampled (i.e. forgotten) at each generation. The sampling procedure for the TOKEN FREQUENCY condition is discussed below.

The BASELINE model represents a frequency-matching grammar with no learning biases.[10] The P-MAP model incorporates a perceptual similarity bias, while the TOKEN FREQUENCY assumes a lexicon where ʈʂ~r forms have higher token frequency than ʈʂ~t

---

[10]Although, note that this model has a smoothing term which serves only to prevent model overfitting. The smoothing term penalizes models with a few closely-fitted constraints, and instead prefers for weight to be more evenly distributed across constraints.

forms. These models are included to confirm that a markedness bias improve model predictions more than alternative accounts which have been substantiated by prior research (Hare & Elman 1995; White 2013, 2017). If reanalysis is in fact driven by a markedness bias in Malagasy, then the Markedness model should outperform the other three models.

Baseline condition (flat-prior). The Baseline model (labeled Base in Table 3.14) assigns each constraint the same μ of 0.

Markedness condition. The Markedness condition (labeled M in Table 3.14) assigns a uniform prior, $\mu = 0$, to all constraints except for the relevant markedness constraint. The markedness constraint *V[-cont,-voice]V is assigned a high prior ($\mu = 5$). This value is higher than the μ assigned to the competing faithfulness constraint *Map(ʈʂ,r), but is otherwise arbitrary. This condition differs from the Baseline condition *only* in the μ value assigned to *V[-cont,-voice]V; the two models are otherwise identical. Note that *r...r] is not given a bias towards high weight despite being a markedness constraint because, as described in §3.3.5, *r...r] does not have support from stem-internal phonotactics.

P-map condition. The p-map condition (labeled P in Table 3.14) has a bias towards higher-weighted faithfulness constraints, scaled by perceptual similarity. The μ of *Map constraints is higher for mappings between perceptually dissimilar sounds, and lower for mappings between perceptually similar sounds. In addition, all markedness constraints are assigned $\mu = 0$.

To approximate perceptual similarity, I adopt White's (2013; 2017) method of using confusability as a measure of perceptual similarity, where the confusability of two speech sounds is determined according to the results of standard identification experiments.[11] As there are no confusability experiments for Malagasy, I use results from Warner et al.

---

[11]Specifically, confusability values are used to train a separate MaxEnt model, whose weights become the priors for the main model. Details of implementation are given in (White 2013, 2017).

(2014), a study of consonant confusability in English, as a proxy. I use Warner et al. (2014) because unlike other studies of English consonant confusability (e.g. Wang & Bilger 1973; Cutler et al. 2004), it includes [ɾ] as a stimulus item. English [ɾ] is used in place of Malagasy <r> [r~ɾ]. Additionally, English does not have a retroflex affricate (except allophonically when [t] precedes [ɹ]), so [tʃ] is used as a substitute for [ʈʂ].

TOKEN FREQUENCY condition. In this condition, constraints are assigned a uniform prior, so the model structure looks in essence identical to the BASELINE condition. However, words are sampled according to token frequency, where the higher the token frequency, the more likely a form is to be remembered.
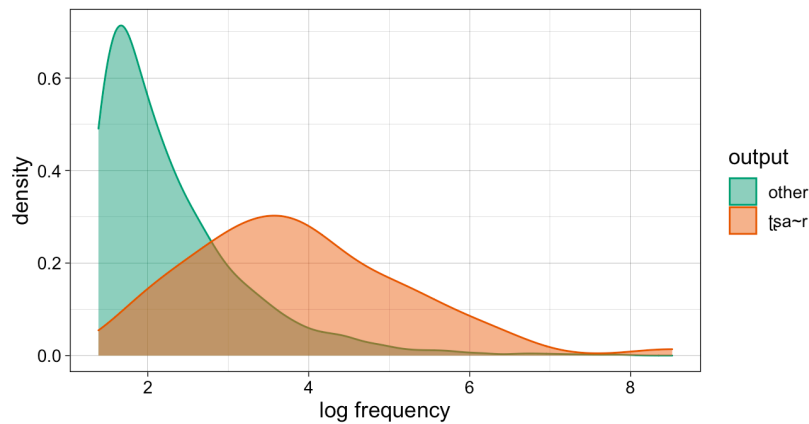
Figure 3.3: Distribution of inputs by token frequency (log)

Token frequency data is not available for Malagasy. Instead, I simulate a lexicon in which ʈʂa~r words are generally more frequent than all other word types. Words are assigned a token frequency on a Zipfian distribution (Zipf 1999). Following this, a log function was applied to raw frequencies. This is to capture the intuition that in acquisition, a difference between high-frequency words is less likely to be relevant (e.g. n = 1000 vs. 1010), while a difference between low-frequency words matters much more (e.g. n = 1 vs. 10) (Marcus et al. 1992; Jackson & Cottrell 1997; Polinsky & van Everbroeck 2003). The token frequency distribution of ʈʂa~r inputs vs. other weak stem types is given in Fig. 3.3.

Constraint	σ^2	BASE	M	P	FREQ
*[ʂ]V	1	0	0	0	0
*k]V	1	0	0	0	0
*n]V	1	0	0	0	0
*r...r]	1	0	0	0	0
*MAP(tr,r)	1	0	0	5.13	0
*MAP(tr,t)	1	0	0	2.82	0
*MAP(n,m)	1	0	0	1.83	0
*MAP(k,f)	1	0	0	2.76	0
*MAP(k,h)	1	0	0	0	0
*V[-cont,-vc]V	1	0	5	0	0

Table 3.14: Constraints and bias terms by condition (P = p-map condition, M = markedness condition, FREQ = Token frequency condition)

3.3.8 Model results

After 50 iterations, the MARKEDNESS model clearly outperforms the other models. This is seen in Table 3.15, which shows the proportion of variance explained (R^2) and log likelihood ($\hat{L}$) for each model after 50 iterations, where the model predictions are fit to the modern Malagasy distribution. Additionally, Fig. 3.4 compares the model fit (R^2) in the four conditions over iterations (again, fit to the modern Malagasy data).

Condition	R^2	($\hat{L}$)
BASELINE	0.60	-9273
P-MAP	0.58	-9618
TOKEN FREQUENCY	0.89	-7233
MARKEDNESS	0.98	-5960

Table 3.15: Results after 50 iterations: Proportion of variance explained (R^2) and log likelihood ($\hat{L}$), of model predictions fit to modern Malagasy

For the two control models (BASELINE and P-MAP), model fit does not improve over iterations ($R^2 \approx 0.6$). The TOKEN FREQUENCY model does better, achieving a higher likelihood and model fit ($R^2 = 0.89$). In comparison, the MARKEDNESS model achieves the highest log-likelihood and is able to account for over 98% of the variation in the observed Malagasy data ($R^2 \geq 0.98$). It does so after around 30 iterations.
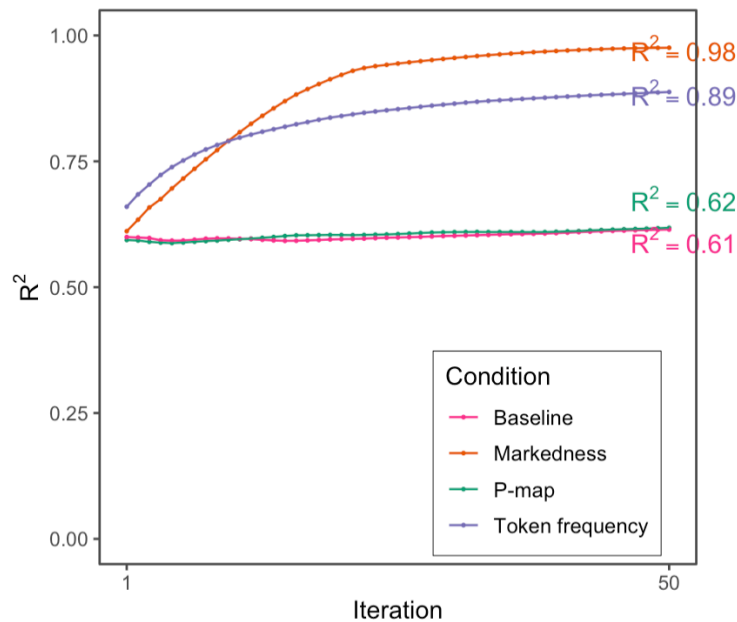
Figure 3.4: Model fit by conditions over 50 iterations (mean of 20 trials)

A more detailed examination of model predictions shows that the bulk of improvement in model fit is driven by changes to ʈʂa-final weak stems. Consider Fig. 3.5, which plots the change in predicted probabilities over model iterations, for ʈʂa-final weak stems. Rates of alternation in the input data (PMP) and modern Malagasy (Mlg) are given at the endpoints of the x-axis for reference. The candidates labeled with "(r...)" have a preceding [r] in the stem; for example, "(r...)ʈʂ∼t" refers to input-output pairs like ['vuriʈʂa]∼[vu'ritana]. Recall that for these words, the observed alternant is exceptionlessly [t]. A model of Malagasy reanalysis must predict reanalysis of t→r, while also preserving the exceptionless r-dissimilation pattern. The non-alternating candidates, 'ʈʂ∼ʈʂ' and '(r...)ʈʂ∼ʈʂ', are omitted for ease of interpretation, since they are assigned zero or near-zero probability by all models.

In the condition with a markedness bias, the model successfully predicts an increase in the ʈʂ∼r alternating candidate (labeled ʈʂ∼r), and almost matches the Malagasy distribution by 50 iterations. At the same time, the model also correctly predicts ʈʂa∼r alternation for candidates with a preceding [r].

88
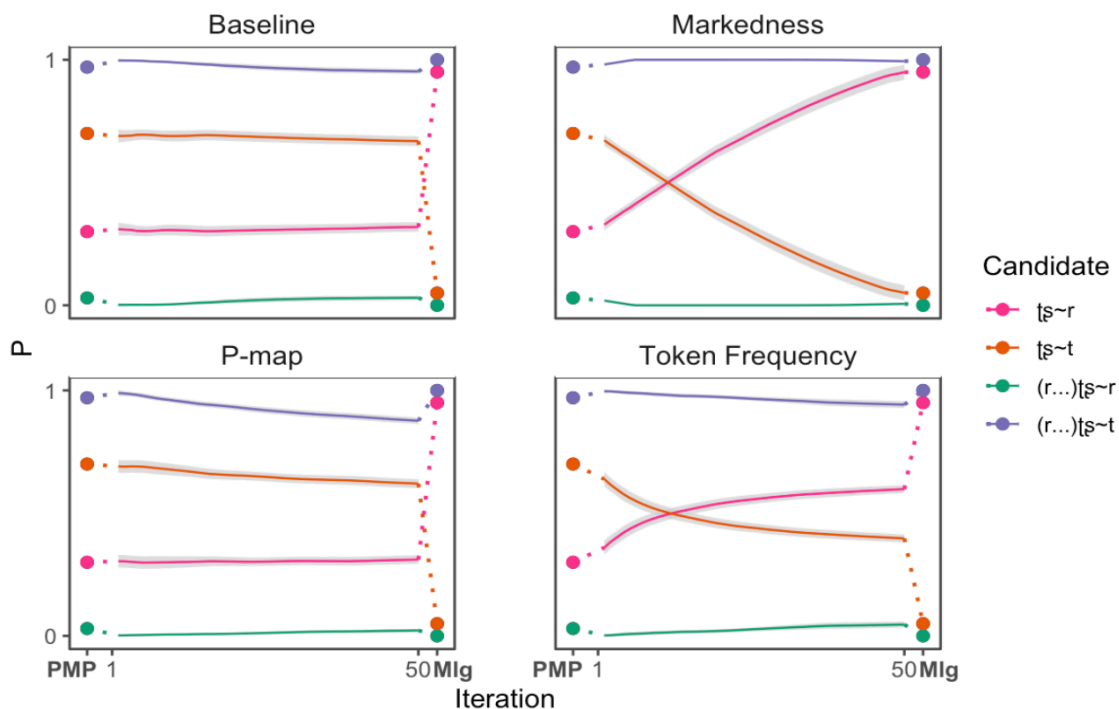
Figure 3.5: Predicted probabilities of candidates over 50 iterations for ʈʂa-final weak stems (mean of 20 trials). Grey intervals indicate standard error, and observed rates of alternation in PMP and Malagasy are given for reference.

The TOKEN FREQUENCY model has inputs which were designed to purposefully favor reanalysis towards /r/ (as r-alternating stems were given overall higher token frequencies). As such, this model actually predicts change in the right direction. However, it predicts reanalysis towards /r/ at a magnitude that is too small to match the modern Malagasy distribution even after 50 iterations.

In fact, even when the input is unrealistically skewed towards giving ʈʂa~r forms higher token frequencies, model predictions do not match the magnitude of reanalysis found in Malagasy. Fig. 3.6 below shows the MARKEDNESS and TOKEN FREQUENCY conditions compared against one extra condition, labeled here as TOKEN FREQUENCY (SKEWED). In this new condition, ʈʂa~r alternating forms have have higher token frequency than all other forms. In addition, ʈʂa~t alternating forms are assigned lower token frequency than all ka- and na-final weak stems (as with before, frequencies are in

89

a Zipfian distribution). We see that although the model with highly skewed token frequencies predicts a greater magnitude of reanalysis towards [r], it still does not perform as well as the MARKEDNESS model.
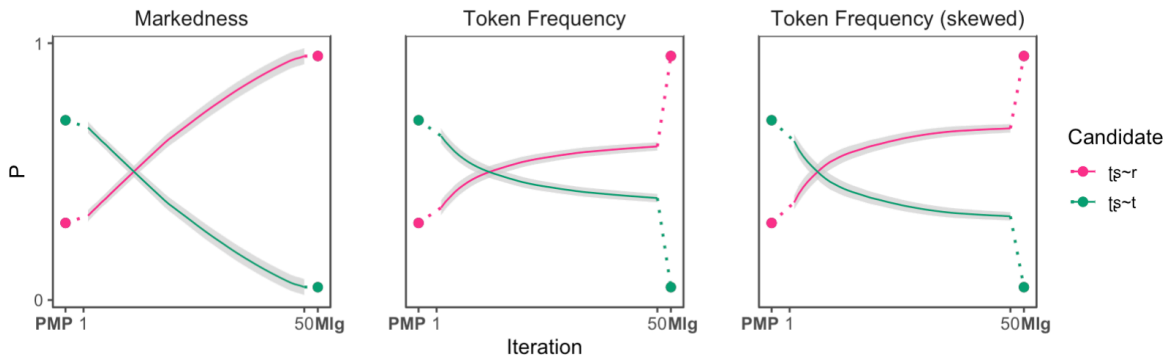
Figure 3.6: Different token frequency distributions vs. markedness condition (50 iterations, mean of 20 trials)

For ka- and na-final weak stems, all four model conditions perform similarly well. This is demonstrated in Fig. 3.7 and Fig. 3.8, which respectively plot model predicted probabilities (over 50 iterations) for ka- and na-final weak stems. All four conditions assign high probabilities to the h-alternating candidate for ka-final weak stems, and the non-alternating candidate for na-final weak stems. These results show that a markedness-biased model is able to predict frequency-matching in environments where markedness is neutral (i.e. where all alternants are equally marked).
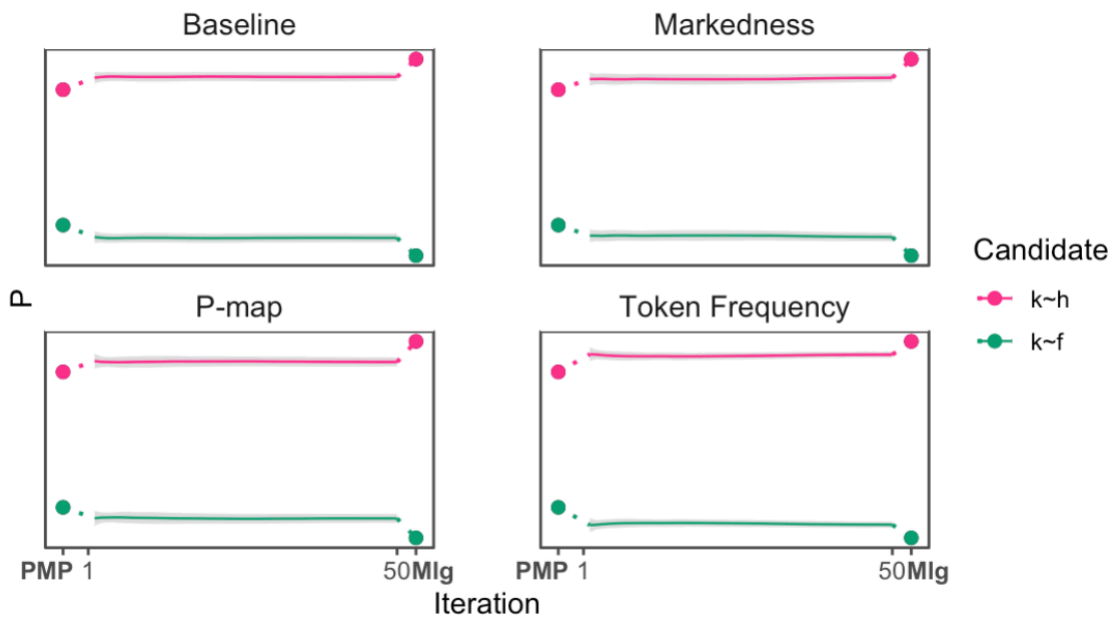
Figure 3.7: Predicted probabilities of candidates over 50 iterations for ka-final weak stems (mean of 20 trials).
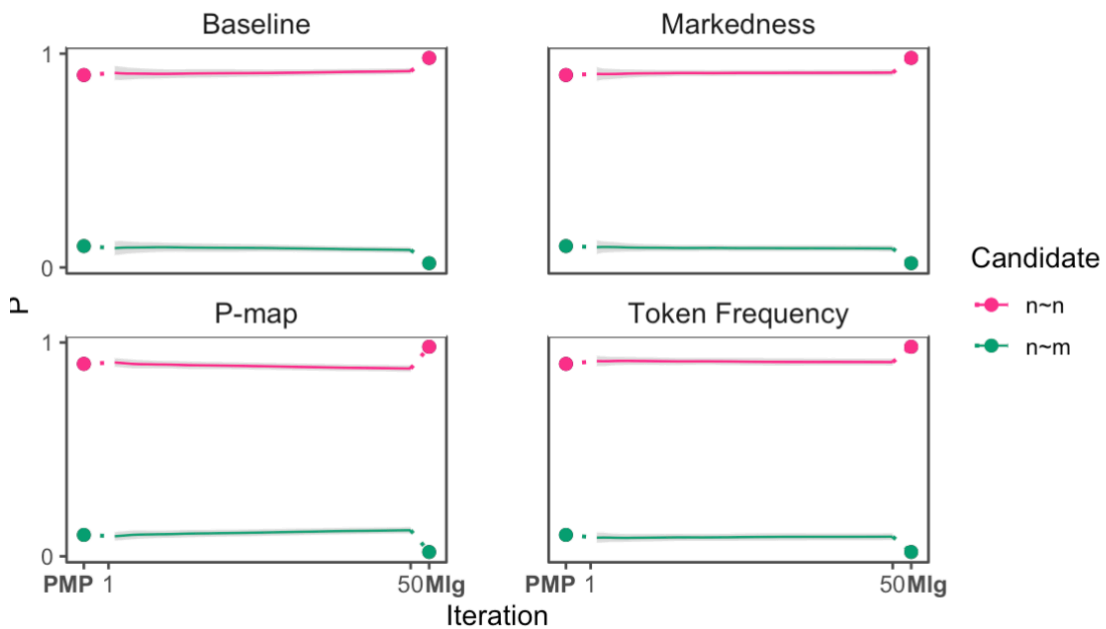
Figure 3.8: Predicted probabilities of candidates over 50 iterations for na-final weak stems (mean of 20 trials).

3.3.9 Iterated learning and dialect divergence

The figures presented so far show model results averaged over multiple trials. When we instead consider individual runs, there is some variation in model predictions. This is demonstrated in Fig. 3.9, which shows predictions of the markedness-biased model for ʈʂa-final weak stems, by individual runs. The dotted grey line, which shows model predictions after 25 iterations, is included as a reference point.

Across all trials, the general trend is reanalysis towards ʈʂa∼r alternation. However, some trials regularize to r-alternation more quickly; for example, Run 18 predicts near-exceptionless rates of ʈʂa∼r alternation after around 25 iterations. On the flip side, some runs also show weaker effects of markedness; for example, Runs 10 and 16 both predict rates of ʈʂa∼r alternation to hover around the 50% mark.

This type of variation over model runs could be taken to reflect individual variation, and more broadly dialect divergence. That is, markedness bias may affect different speakers to a different degree, causing the same language to undergo dialect divergence.

3.4   Chapter conclusion

Overall, comparison of PMP and modern Malagasy suggests that reanalysis of ʈʂa-final weak stems has occurred in a way that is not predicted by frequency-matching approaches. Instead, I proposed that reanalysis in the direction of t→r is motivated by a markedness constraint against intervocalic stops. In §3.3, this proposal is confirmed by comparing a frequency-matching model of reanalysis against one with a markedness bias. More concretely, reanalysis of the ʈʂa-final weak stems is not predicted by distributional information within a paradigm, but can be modeled as the effect of a markedness constraint against intervocalic (voiceless) stops. The results for Malagasy are also consistent with the principle of active markedness, as constraints against intervocalic stops were found in a phonotactic grammar of Malagasy stems.
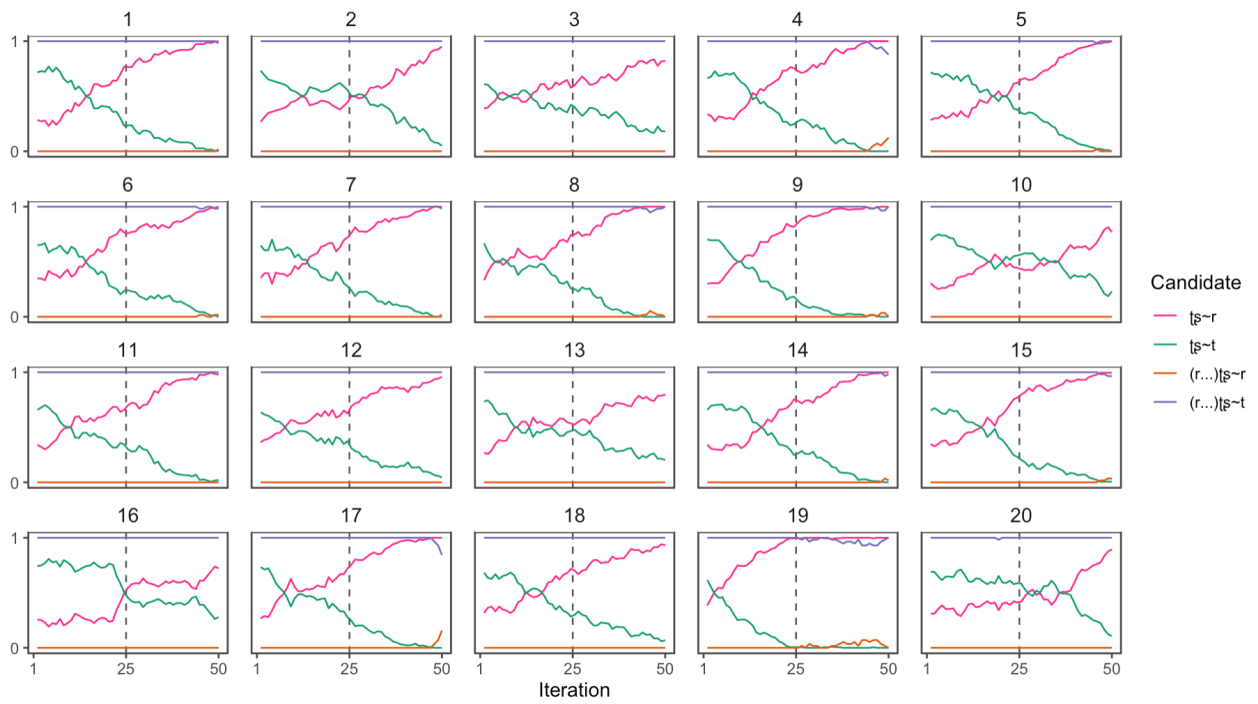
Figure 3.9: Predictions of markedness-biased model by individual runs (ʈʂa-final weak stems)

CHAPTER 4

Case study 2: Samoan thematic consonants

Proto-Oceanic (POc) is the reconstructed ancestor of the Oceanic subgroup of the Aus-
tronesian language family. POc allowed final consonants, but many Oceanic languages
(including Central Pacific, Western Oceanic, and South-East Solomonic) regularly lost
them. In particular, the entire sub-family of Polynesian languages underwent final con-
sonant loss.[1] For these languages, consonants were maintained in suffixed forms, re-
sulting in $\emptyset$∼C alternations, where a consonant of unpredictable quality surfaces under
suffixation (e.g. POc *inum/*inum-ia → Samoan *inu/inu-mia* 'to drink' and POc *suat/suat-
ia→sua/sua-tia*). These consonants are traditionally called thematic consonants. The table
in (36) gives examples of $\emptyset$∼C alternations in Samoan using the ergative suffix; these will
be discussed in more detail below.

(36)   *Examples of Samoan thematic consonant alternations*

STEM	ERGATIVE	POc	GLOSS
pulu	pulutia	*bulut	'to plug up'
laka	lakasia	*lakas	'to step over'
tutu	tuŋia	*tutuŋ	'to light a fire'
fulu	fulu-a	*pulu	'to rub, wash'

At this point, it should be noted that the consonant which surfaces under suffixation
is often a reflex of a historical final consonant. When the historical stem is vowel-final

---

[1]While Malagasy (Chapter 2) and the Polynesian languages both underwent sound changes which re-
moved codas, Malagasy achieved this by final vowel epenthesis, while the Polynesian languages achieved
this by final consonant deletion.

(or ends in a consonant that has been deleted in all environments due to regular sound change), no thematic consonant should surface (e.g. *fulu/fulu-a*, <POc *pulu).

However, the thematic consonant that surfaces under suffixation often does not correspond to the historical final consonant, indicating that reanalysis has occurred. In many cases, languages with thematic consonant alternations have leveled towards a few relatively predictable "default" consonants. In some languages like Hawaiian, consonant alternations have been completely lost and there is just one consonant that surfaces (/ʔ/ for Hawaiian).

Previous works on reanalysis in thematic consonant alternations have opted for a frequency-driven approach. For example, Blevins (2008) finds that where leveling has occurred, it is inconsistent with phonological markedness predictions. Instead, reanalysis tends to be towards the more frequent allomorphs (e.g. towards /tia, ia, a/ in Māori; Hale 1973; Blevins 1994). Where there are exceptions to this pattern, reanalysis is based on factors such as semantics, rather than phonological markedness (e.g. extension of /m/ as the thematic consonant for kinship terms in Manam; Lichtenberk 2001).

However, I argue that a more nuanced view is needed, in which reanalysis can be affected by both frequency and markedness effects. In this chapter and the following chapter, I look at reanalyses of thematic consonant alternations in two languages, Samoan and Māori, and find that both show interacting effects of frequency and markedness. Among the Polynesian languages with ∅~C alternations, Māori and Samoan are particularly informative case studies because both are relatively conservative in maintaining the final consonant contrasts present in Proto-Oceanic. They are therefore suitable for studying intermediate stages of reanalysis (vs. languages like Hawaiian).

This chapter focuses on Samoan. As a preview, Samoan reanalysis is largely towards the more frequent alternants, but in a way that is modulated by markedness effects. Specifically, I find that reanalysis in Samoan is sensitive to transvocalic consonant co-occurrence restrictions (i.e. OCP-place effects; McCarthy 1994). Findings are also consistent with my proposal that markedness effects are restricted to those already present in

stem phonotactics.

The rest of this chapter will be organized as follows. §4.1 will give an overview of Samoan phonology and thematic consonant alternations. While the development of weak stem alternations is relatively straightforward and shared by all Polynesian languages, I will also discuss it briefly in the context of Samoan. In §4.2, I provide evidence that transvocalic consonant co-occurrence restrictions are present in both Samoan stem phonotactics and Proto-Polynesian (which represents an earlier stage of Samoan). §4.3 discusses patterns of reanalysis in Samoan, with a focus on environments where there is a mismatch between historical distributions and the modern Samoan distribution. Finally, in §4.4, I demonstrate that a model of reanalysis which incorporates OCP-place effects outperforms one that is purely frequency-matching.

4.1 Background: Samoan phonology

Samoan is an Oceanic language of the Nuclear Polynesian sub-branch, spoken primarily in the Independent State of Samoa and the United States Territory of American Samoa, with about 370,000 speakers in all countries (Eberhard et al. 2023). There is a sizeable population of speakers living in New Zealand, Hawaii, the United States West Coast, and Australia.

Samoan can broadly be split into two dialects, respectively Western Samoan and American Samoan. The data from this paper (and most linguistic research on Samoan) is from Western Samoan. Though no systematic study has been done, scholars generally agree that within Western Samoan, the degree of geographical dialect variation is small (Mosel & Hovdhaugen 1992).

Samoan has two registers of speech, respectively *tautala lelei* (literary language) and *tautala leaga* (colloquial and traditional oratory language). The two registers differ primarily in that *tautala lelei* preserves more segmental contrasts (discussed below), but native speakers are generally cognizant of both levels. All data in this paper are from the

tautala lelei register, as it is the register described in dictionaries and the subject of most scholarly work on Samoan.

Samoan is relatively well-documented. Linguistic descriptions of Samoan date back to missionary texts from the 1800s. Since then, there has been extensive descriptive work, including grammars (e.g. Churchward 1951; Mosel & Hovdhaugen 1992) and dictionaries (e.g. Pratt 1862/1893; Violette 1880; Milner 1966). Formal analyses of Samoan have primarily covered syntax and morphosyntax (e.g. Pawley 1962, 1966; Chung 1978; Cook 1988). Work on Samoan phonology is less extensive, but includes Zuraw et al. (2014) on prosody and Alderete & Bradshaw (2013) on stem phonotactics. Moore-Cantwell (2008) was the first work to look at Samoan thematic consonants in detail.

Additionally, the historical subgrouping of Polynesian languages, including Samoan, has been worked on in detail (e.g. Dempwolff 1929; Pawley 1966, 1967; Clark 1973; Greenhill & Clark 2011). Hovdhaugen et al. (1986) has also worked on the sound changes specific to Samoan. Historical comparative data is available in both the Austronesian Comparative Dictionary (ACD; Blust & Trussel 2010) and the Polynesian Lexicon Project (POLLEX; Greenhill & Clark 2011).

In the rest of this section, I give an overview of Samoan phonology and details about the historical development of thematic consonants in Samoan. Unless otherwise noted, descriptive generalizations are taken from Mosel & Hovdhaugen (1992).

4.1.1   Phoneme inventory and phonotactics

Samoan syllables follow a (C)V(V) structure; no codas or consonant clusters are allowed and onsets are optional. Stress is non-contrastive, falling on final long vowels and otherwise on the penultimate vowel (i.e. moraic trochee at the right edge of the word, Zuraw et al. 2014). Note that suffixation can also shift word stress, but this relationship is dependent on suffix size (Zuraw et al. 2014).

Samoan has five vowels /a, e, i, o, u/, all of which also show a two-way length con-

trast. Vowel-vowel sequences are allowed, both in hiatus and as diphthongs. I follow Zuraw et al. (2014) in assuming that /ai, au, ei, ou/ are diphthongs; in stress assignment, diphthongs behave like long vowels, suggesting that they are distinct from other vowel-vowel sequences. Note that Mosel & Hovdhaugen (1992) propose a larger set of diphthongs that additionally includes /eo, oi, ui/.

The consonant inventory (of the *tautala lelei* register) is given in (37). /ʔ/ is phonemic, but described by Mosel & Hovdhaugen (1992) as being "unstable in initial position...elided except in very careful speech". The phonemes given in parentheses (/k, r, h/) are all found only in loanwords or interjections, and not in native words (i.e. words inherited from POc). Additionally, /r/ is often realized as [l] even in careful speech.

(37)　*Samoan consonant inventory (tautala lelei)*

LABIAL	ALVEOLAR	VELAR	GLOTTAL
p	t	(k)	ʔ
f v	s		(h)
m	n	ŋ	
	l (r)		

As discussed above, results will focus on the *tautala lelei* register (in contrast to the colloquial *tautala leaga* register). The phonological differences between the two registers are given in (38). Essentially, *tautala leaga* has three phonological mergers; note that of these mergers, two involve loanword phonemes /k/ and /r/, where /r/ is very unstable in both registers.

(38)　*Mergers in tautala leaga* (Mosel & Hovdhaugen 1992, p. 24)

tautala lelei		*tautala leaga*	Example
/t,k/	→	/k/	/tiː/ "tea", /kiː/ "key" →/kiː/
/n,ŋ/	→	/ŋ/	/ana/ "conduct", /aŋa/ "cave"→/aŋa/
/r,l/	→	/l/	/taro/ "taro", /tatalo/ "prayer"→ /(ta)talo/

4.1.2   Samoan ∅∼C alternations and their historical development

In Samoan, ∅∼C alternations are observed in a variety of suffixal contexts, listed in (39). Where thematic consonants surface in the examples, they are shown in boldface. Note that the suffix /-(C)i/ is non-productive and doesn't have a clearly defined function; Mosel & Hovdhaugen (1992, p. 205) describe it as being used to derive words "with a more specific and narrow meaning". For the rest of this chapter, I focus on just the **ergative suffix**, since it is the most productive of the suffixes in (39).

(39)   *Samoan suffixes with thematic consonant alternations*

FUNCTION	ALLOMORPHS	EXAMPLES
nominalizer	-ŋa, (C)aŋa	tafe/tafeŋa 'to flow/current'
		inu/inu**m**aŋa 'drink/draught'
ergative	-a, -ina, -(C)ia, -na	ʔini/ʔini**t**ia 'to pinch'
		fuli/fuli**s**ia 'to turn, roll over'
Lex	=(C)i	sua/sua**t**i 'uproot/dig up (violently)'
		sulu/sului "dried banana-leaf"
intensifier	-(C)aʔi	oso/oso**f**aʔi "jump/attack"
		ufi/ufiaʔi "cover/covered"

The ergative suffix has the allomorphs /-a/, /-ina/, /-ia/, /-Cia/, and /-na/, where 'C' can be one of consonants /f, m, t, s, l, ŋ, ʔ/. Examples of each allomorph are given in Table 4.1. The two vowel-initial allomorphs /-a/ and /-ina/ are the most frequent, and typically analyzed as the 'default' ergative allomorph. They are productively applied to derived words and loans, and the relative distribution of /-a/ and /-ina/ are decided by a combination of phonological factors (discussed in §4.1.3). Note that /-ia/, which is also a vowel-initial allomorph, is non-productive and relatively infrequent in Samoan. The /-Cia/ allomorphs (i.e. ones which preserve a thematic consonant) are also relatively infrequent and non-productive. Nevertheless, they still account for a substantial proportion of the lexicon; out of 527 stems that take the ergative suffix in Milner (1966), 34% (n = 179/527) have a thematic consonant.

ERG.	STEM	SUFFIXED	GLOSS	POc
a	rere	rere-a	to take	*rere
ia	nofo	nofo-ia	to take	*nofo
ina	iloa	iloa-ina	to see, perceive	*qilo
sia	laka	laka-**s**ia	to step over	*laka**s**
tia	pulu	pulu-**t**ia	to plug up	*bulu**t**
ŋia	tutu	tu-**ŋ**ia	to light a fire	*tutu**ŋ**
fia	utu	utu-**f**ia	to draw water	*qutu**p**
mia	inu	inu-**m**ia	to drink	*inu**m**
lia	tautau	tautau-**l**ia	to hang up	*sau**r**
na	ʔai	ʔai-**n**a	to eat	*kae**n**
ʔia	momo	momo-**ʔ**ia	to break in pieces	*mekme**k**

Table 4.1: Samoan ∅/C alternations

As discussed above, thematic consonant alternations developed from the deletion of final consonants, a process which affected all languages in the Polynesian family. For example, POc *inum/inum-a became Proto-Polynesian *inu/inu-mia following the deletion of final *m. For words that have not undergone reanalysis, the suffix allomorph that surfaces should correspond to the final consonant in POc. This is demonstrated in Table 4.1, where each stem is also given with its POc protoform.

Assuming no reanalyses, stems that historically ended in vowels (and in some consonants, as summarized in Table 4.2 below) should take the vowel-initial suffixes /-a/, /-ia/, or /-ina/. Otherwise, the suffix that surfaces is of the form /-Cia/, where /C/ corresponds to are given with their corresponding forms. As a caveat, when the stem historically ended in *n, the suffix that surfaces is either /-ina/ or /-na/, where /-ina/ surfaces after [a]-final stems, and /-na/ surfaces elsewhere. Note that /-ina/ is homophonous with the default allomorph.

The Samoan ergative suffix descends from the Proto-Polynesian CIA suffix. Historical work suggests that allomorphy of this CIA suffix (between /-a/, /-ina/, /-na/, /-ia/, and /-Cia/) can be explained by regular sound changes between POc and Proto-Polynesian (PPn). These changes are summarized in (40); the reader is referred to Pawley (2001) for an overview discussion. In POc, the ergative was originally two suffixes: the short

transitive marker *-i followed by the third person pronominal clitic, which was variably realized as *-a or *-na (e.g. POc *kila-i-a 'know it' vs. *POc *tabu-i-na 'to be banned'). Over time, the suffixes *-a and *-na ceased to be productive, resulting in the development of the Samoan ergative suffixes /-ia/ and /-ina/. Although the consensus is that *-i and *-a/na were separate suffixes in POc, there is some debate about the historic meaning of the second *-a/na suffix. I follow Churchward (1951) and Pawley (1966) in calling the POc *-a/na suffix a pronominal marker, but alternate hypotheses are discussed in Pawley (1966).

(40)   *Development of the Samoan ergative suffix*

Suff.	/ia/	/ina/	/a/	/Cia/	/ina/	/na/	
POc	*pana-i-a	*tabu-i-na	*tari-i-a	*puat-i-a	*pulan-i-a	*talun-i-a	CHANGE
	-	-	tali-a	-	-	-	*i-deletion[A]
	-	-		-	pula-ina	talu-ina	Metathesis[B]
	-	-		-		talu-na	ina→na
Sam.	fana-ia	tapu-ina	tali-a	fua-tia	pula-ina	talu-na	
	'sit'	'shoot'	'wait'	'bear fruit'	'be bright'	'forest'	

A. *i-deletion: *i is deleted after *i,*e. **Note:** Evans (2001) argues that deletion of *i happens after stems ending in all vowels other than *a, but Pawley (1962) proposes that the evidence for deletion after *o and *u is less conclusive, citing forms like [nofo-ia] (<*nofo) 'to sit', where *i is maintained after the back vowels.

B. Metathesis of *nia to *ina after *a

Additionally, the POc transitive *-i had phonologically conditioned allomorphy, and was deleted after stems ending in *i and *e (e.g. POc *kani-∅-a 'eat it' vs. *kila-i-a 'know it'). Evans (2001) actually proposes that *i deleted after all non-*a vowels, but Pawley (2001) suggests that the evidence for deletion after *o and *u is less strong. Either way, *i-deletion derives the /a/ ergative allomorph.

Following *n-final POc words, the observed suffix allomorphs are /-ina/ and /-na/, rather than /nia/. /-ina/ arose by metathesis from pre-Polynesian *-nia, mainly when the verb base ended in *a. /-na/ likely also arose from metathesis of *-nia, followed

by deletion of the *i vowel (e.g. POC *aŋin-ia > *aŋiina > Samoan [aŋi-na] 'blow'). Although, it should be noted that metathesis of *-nia was less consistent when the verb did not end in *a; for these words, /nia/ is the observed reflex in other Polynesian languages (e.g. Fijian, Niuean, Hawaiian).

Based on the historical developments described in (40), /-ia/ and /-a/ should have a phonologically predictable distribution, with /-a/ surfacing generally, and /-ia/ surfacing after [a]-final stems (and occasionally after [o]- and [u]-final stems). In Samoan, however, /-ina/ has generally been extended to the environments where /-ia/ would be expected to surface.

Finally, Table 4.2 summarizes the regular sound correspondences between POc and Samoan. Based on these, we can infer the ergative allomorph that should surface if no reanalysis had taken place.

POC	PPn	Sam	Ergative
*p,*pw	*f	f	fia
*t, *j, *d	*t	t	tia
*l, *r, *dr	*l, *r	l	lia
*k, *g	*k	ʔ	ʔia
*m	*m	m	mia
*n, *ñ	*n	n	na, ina
*ŋ, *mw	*ŋ	ŋ	ŋia
*s	*s	s	sia[2]
*c	*h	∅	(in)a
*q	*ʔ	∅	(in)a
*y, *R	∅	∅	(in)a

[1]POc also had phonemes *b/*bw and *w, whose Samoan reflexes are [p] and [v]. These are excluded here because they are not found word-finally in POc, and therefore never reflect as thematic consonants. [2]POc *s became [s] in Futunan and Samoan, but [h] in other Polynesian languages.

Table 4.2: Samoan reflexes of POc final consonants[1]

4.1.3   Distribution of /-a/ and /-ina/

Throughout the rest of this chapter, I refer to both /-a/ and /-ina/ as the 'default' ergative allomorph, originating from historically vowel-final stems. However, it should be noted that in Samoan, the relative distribution of /-a/ and /-ina/ is partially predictable from several factors: stem-final vowel, mora count, and prosodic shape. Although a complete analysis of the phonological conditioning of /-a/ and /-ina/ is not the focus of this chapter, this section will give an overview of these factors.

The data used here is 362 stem-ergative pairs taken from Milner (1966) and Pratt (1862/1893). Note that the ergative forms listed in Milner (1966) can differ from the stem in several ways, detailed in (41). First, as seen in (41a), some words undergo vowel shortening to avoid feet consisting of a long-vowel syllable plus a light syllable (Zuraw et al. 2014). This process of *trochaic shortening* (Hayes 1995) can result in vowel length alternations between stem and suffixed forms.[2] Stems can also be reduplicated in the unsuffixed form, but not when suffixed as in (41b). The opposite can also be true, as in (41c). Finally, as demonstrated by (41d), the ergative may take a suffix, most commonly the causative /faʔa-/. In all of these cases, I code the prosodic shape and mora count based the base of suffixation in the ergative; for example, [faʔapito-a] is coded as having a 4-mora base.

(41)   *Stem-ergative pairs*

	stem	ergative	gloss	
a.	tusi	tuːsi-a	'to write'	(trochaic shortening)
b.	itiiti	iti-a	'little, reduced'	(reduplication in stem)
c.	eto	etoeto-ina	'to lick'	(reduplication in suffixed form)
d.	pito	faʔapito-a	'to be next'	(prefix /faʔa-/)

---

[2]In Samoan, trochaic shortening is lexically specific; Milner (1966) lists both roots that always have a short vowel (e.g. [fusi], [fusi-a] 'to hug'), roots that alternate (e.g. [tusi, tuːsi-a] 'to write'), and words that always have a long vowel like [paːsi, paːsi-a] 'to be tired of'. Note that more recent descriptions find that words like [paːsi] are shortened to [pasi] (Mosel & Hovdhaugen 1992), so the troachic shortening pattern could be regularizing.

The first tendency in the relative distribution of /-a/ and /-ina/ is quantity-sensitivity, such that longer words prefer /-ina/. This is seen in Fig. 4.1, which plots the relative frequency of /-a/ and /-ina/ by the mora count of the base. In words with five or more moras, the suffix is is exceptionlessly /-ina/.
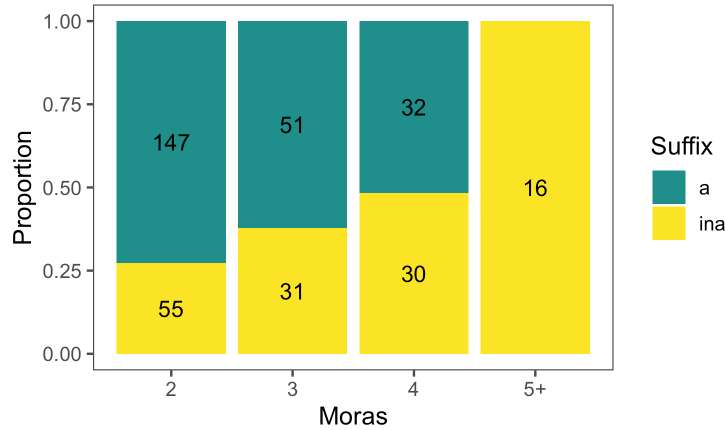
Figure 4.1: Distribution of /a/ and /ina/ by number of moras in base
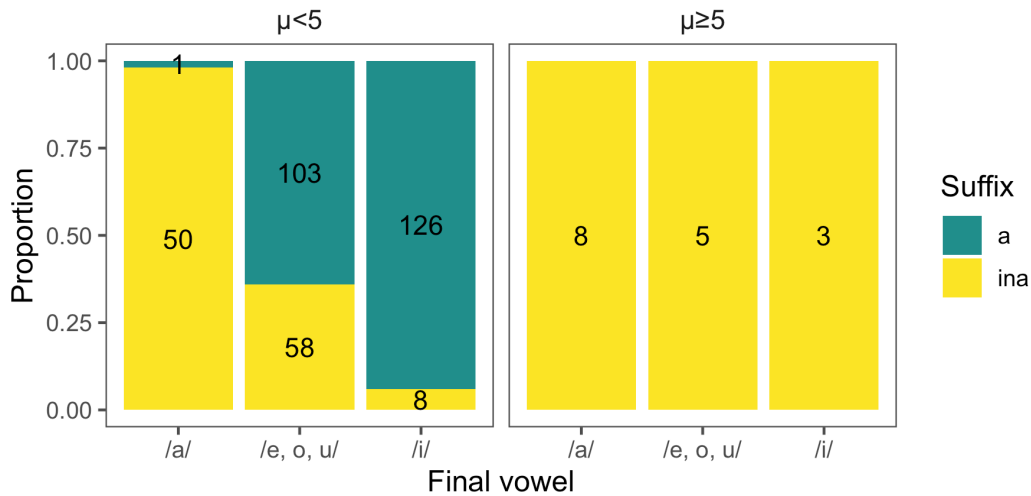
Figure 4.2: Distribution of /-a/ and /-ina/ by stem-final vowel (<5 moras)

Within words that are less than five moras, the relative preference for /-a/ or /-ina/ is correlated with identity of the final vowel and prosodic shape. As seen in Fig. 4.2, for these shorter words, /-a/ almost never follows [a]-final stems, while /-ina/ almost never

follows [i]-final stems. This pattern is near-exceptionless, but there are a few exceptions (e.g. [asa-a] 'wave.ERG', [tipi-ina] 'cut.ERG'). This vowel conditioning effect can be explained as the combined pressures of OCP-vowel across morpheme boundaries (e.g. [asa-a] violates *a-a) and a constraint requiring that morphemes have some unique output exponent (e.g. [asa:] would violate such a constraint). de Lacy (2002) adopts a similar analysis for Māori, where passive suffix allomorphy is sensitive to similar constraints.

Finally, prosodic shape also matters; Fig. 4.3 shows the distribution of /-a/ and /-ina/ by prosodic shape. This figure includes only the subset of forms that are under five moras and end in [e, o, u]; other stems are excluded because, for reasons discussed above, they have near-categorical behavior. H(eavy) refers to (C)V: syllables, L(ight) refers to (C)V syllables, and D(iphthong) refers to (C)VV syllables (where VV is one of /ai, au, ei, ou/).

The patterns that emerge are as follows: in stems of the shape D, LL, DLL, or HL, the preferred suffix allomorph is /-a/. Following H and LLL stems, the preferred suffix allomorph is /-ina/ For the LLLL stems, there is a roughly even split between /-a/ and /-ina/.
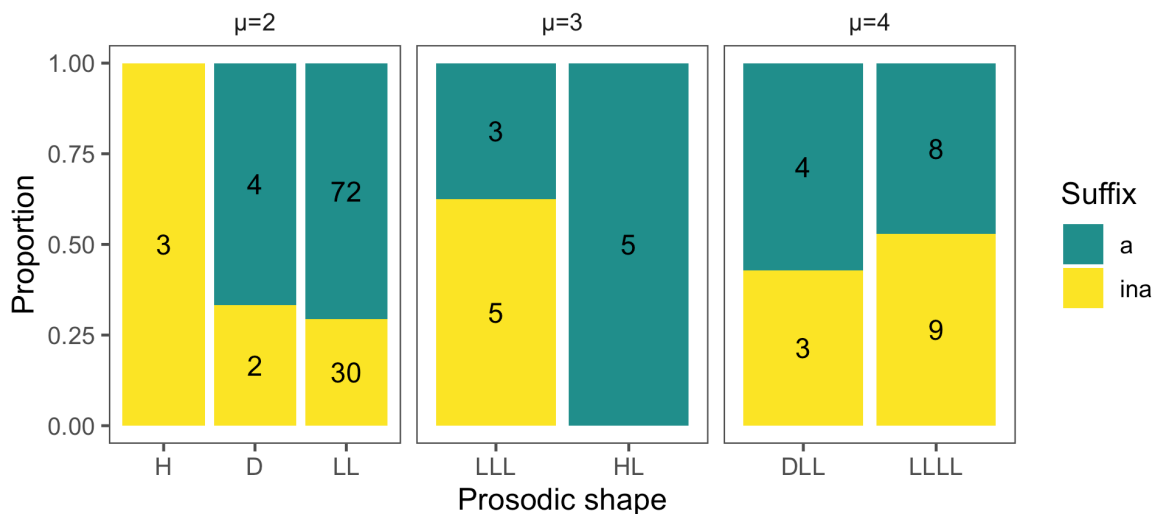
Figure 4.3: Distribution of /-a/ and /-ina/ by word-prosody (<5 moras, final vowels /e,u,o/)

H = (C)V:, L = (C)V, D = (C)VV (where VV forms a licit diphthong)

4.2 Consonant OCP in stem phonotactics

In this section, I provide data showing that Samoan has a gradient dispreference for transvocalic homorganic consonants. In other words, there are effects of the Obligatory Contour Principle on place (OCP-place). Gradient OCP-place effects are well attested in the literature. These effects were first noted in modern linguistics by Greenberg (1950) and McCarthy (1988, 1994) for Arabic, and have since been substantiated by several empirical case studies, including: Muna (Coetzee & Pater 2008a,b), English (Berkley 1994, 2000a), Tigrinya (Buckley 1997), Japanese (Kawahara et al. 2006), and Chol (Gallagher & Coon 2009).

OCP effects have also been documented across multiple Polynesian languages (Krupa 1966, 1967, 1971). Alderete & Bradshaw (2013) conduct a detailed and comprehensive quantitative study of Samoan phonotactics and find gradient OCP-place effects. In particular, they find near-exceptionless OCP-place restrictions for labials (/p, f, v, m/, penalizing words such as *[fuma]). They also find a strong OCP-place effect for coronals that is sensitive to manner, such that OCP-place effects are stronger for coronals which share the same manner of articulation (e.g. *[nula] is marked because [n] and [l] are both coronal sonorants).

Alderete and Bradshaw's results, while quite comprehensive, run into two potential methodological issues. First, they use Observed/Expected (O/E) as a metric for quantifying phonotactically over- or under-represented sequences. However, Wilson & Obdeyn (2009) show that this method is problematic because it cannot control for the confounding effect of interacting constraints. Additionally, Alderete and Bradshaw discuss but do not control for effects of pseudoreduplicants (i.e. forms like [lalaŋa] 'plait, weave', which may originally have been reduplicated but are now fossilized as a monomorphemic word). Reduplicated forms often do not adhere to the same phonotactic restrictions as other roots (Hayes & Jo 2020). Relatedly, there is a cross-linguistic tendency for identical syllables to be preferred over other syllables (aggressive reduplication; Zuraw 2002). It's

therefore possible that forms like [lala] are better than [lila] even though both violate coronal sonorant OCP; this could in turn obscure the effects of OCP-place for identical segments.

With these points in mind, this section will confirm Alderete and Bradshaw's results using two datasets: one that is taken direction from Alderete & Bradshaw (2013) and one without sequences of identical syllables (e.g. [papa], [ʔoʔo], [fafano]). In addition, I adopt a MaxEnt phonotactic grammar (Hayes & Wilson 2008; Wilson & Obdeyn 2009) instead of the O/E method.

4.2.1 Data and basic pattern

The data I use is taken from Alderete & Bradshaw (2013). Their list contains monomorphemic headwords from the Milner (1966) dictionary (i.e. unbound roots). Loanwords and classificatory names (of animals, seafood, plants, etc.) were excluded, resulting in a list of 1640 roots. I also compile a separate list with (pseudo-)reduplicated forms excluded; this list has 1,498 roots.

Figures 4.4 shows counts of all transvocalic consonant-consonant sequences (i.e. C_1VC_2) in these data, where long vowels and diphthongs also count as an intervening V. Fig. 4.4a uses Alderete and Bradshaw's original root list, while Fig. 4.4b uses the smaller list with reduplicants removed. C1-C2 combinations that never occur are labeled '0', and frequent ones (n>25) are labeled with their counts.

Qualitatively, we can observe trends consistent with those found by Alderete & Bradshaw (2013); there is a strong dispreference for labial-labial sequences and a dispreference for coronal-coronal sequences which share the same manner of articulation (e.g. [s...t], [n...l]). In 4.4a, OCP effects appear to hold for similar, but not identical segments. For example, [v...p] is never attested, but [v...v] is relatively well-attested. However, this effect is much weaker once we exclude reduplicated syllables, as seen in 4.4b. In general, along the diagonal outlined in 4.4b, C1-C2 co-occurrences tend to be less frequent.

107

(a) All roots (n=1640) (b) No reduplicated forms (n=1498)
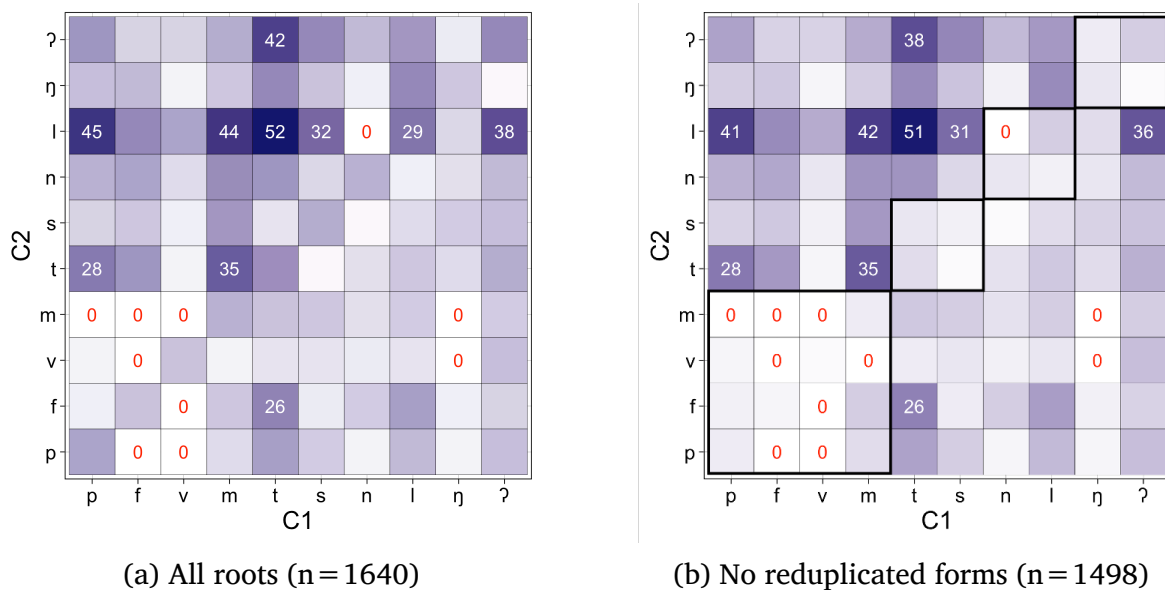
Figure 4.4: Consonant-consonant co-occurrences in Samoan

Additionally, [ŋ...m] and [ŋ...v] are never attested. This could be an accidental gap, and also in part be due to the fact that across Polynesian languages, labials are preferred in initial syllables while dorsal consonants are preferred in non-initial syllables (Krupa 1966).

4.2.2 A statistical model of OCP-place effects

Following Wilson & Obdeyn (2009), OCP-place effects are confirmed using a MaxEnt phonotactic grammar. This method allows for constraint interaction and can therefore control for the lexicon's baseline (dis)preferences for each consonant (See Wilson & Obdeyn 2009 for a more in-depth discussion of the benefits of MaxEnt relative to the O/E method).

The input was all C1-C2 sequences extracted from the list of roots (after removing reduplicants). The constraint set includes singleton constraints penalizing each conso-nant in each position (e.g. *p/C1, *p/C2, *f/C1, *f/C2, etc.); these are used to control for the baseline frequency of each consonant. Additionally, I tested the OCP-place constraints listed in (42). Note that this phonotactic model has a relatively simplified input set and

constraint set, meant to clearly demonstrate the effects of OCP-place in Samoan stem phonotactics. In my modeling results in subsequent sections, I also construct a more nuanced phonotactic grammar using the UCLA Phonotactic Learner (Hayes & Wilson 2008).

(42)   *OCP constraints*

Constraint	Example violations
OCP-place	pama, tasa, nata
OCP-LAB	pama, pava, vama
OCP-LAB-SON	mama, papa, pafa
OCP-COR	tasa, tasa, tala
OCP-COR-SON	nala, lala, tasa
OCP-BACK	ŋaʔa, ʔaʔa, ŋaŋa
OCP-BACK-SON	ŋaŋa, ʔaʔa

The literature on OCP-place shows that crosslinguistically, OCP-place restrictions do not apply with equal strength to all sequences of homorganic consonants. Instead, there is often a stronger effect of OCP-place when two segments agree on one of more of a set of non-place features, referred to in the literature as *subsidiary features* (McCarthy 1988; Yip 1989; Padgett 1991a,b; Wilson & Obdeyn 2009). As such, I also tested for the effect of subsidiary features.

Following Coetzee & Pater (2008a), I implement multiple OCP-place constraints, each a combination of place (LABIAL, CORONAL, BACK) and subsidiary features (SONORANT, CONTINUANT, NASAL, VOICE). For example, OCP-COR-SON penalizes all homorganic sequences of coronals that also agree in [sonorant]; this includes sequences like [n...l] (where both segments are [+sonorant] and sequences like [t...s] (where both segments are [-sonorant]. The constraint set show in in (6) is narrowed down from this larger set of OCP-place constraints.

Note that /ʔ/ was historically *k, and that sound change of *k>ʔ occurred relatively recently, sometime between Proto-Polynesian and modern Samoan. Perhaps because of this, phonologically, /ʔ/ still patterns like a dorsal consonant. More importantly, /ʔ/ was

still conceivably realized as [k] during at least part of the reanalyses that resulted in the modern-day pattern. For these reasons, in my model implementation, I treat /ʔ/ and /ŋ/ as belonging to the same natural class, captured under the place feature BACK. Therefore, OCP-BACK penalizes sequences like [ŋ...ʔ].

Table 4.3 shows the constraint weights found by the model for each OCP-place constraint; constraints were tested for significance using the Likelihood Ratio Test, by comparing a maximal model (with all constraints included) against one with the target constraint excluded (Hayes et al. 2012). In the table, ΔL shows the improvement in log-likelihood from adding the target constraint (a larger positive value indicates greater improvement in model fit).

Constraint	w	ΔL	p
OCP-place	0.14	-0.01	n.s.
OCP-LAB	0.88	6.03	0.0005***
OCP-LAB-SON	0.70	1.54	n.s. (0.08)
OCP-COR	0.00	-0.01	n.s
OCP-COR-SON	1.50	34.76	7.56×10^{-17}***
OCP-BACK	1.03	3.94	0.002**
OCP-BACK-SON	0.00	0.01	n.s.

Table 4.3: OCP constraint weights learned by the phonotactic model

Overall, I replicate Alderete and Bradshaw's (2013) findings, and results are consistent with the qualitative observations from Fig. 4.4. First, the model learned a significant weight for OCP-LAB, showing a general dispreference for homorganic labials, regardless of manner of articulation. OCP-LAB-SON is non-significant, suggesting that for labials, subsidiary features have less influence.

For coronals, the general constraint OCP-COR is non-significant, but OCP-COR-SON is actually strongly significant; it has the highest weight out of the constraints tested, and led to the biggest improvement in log-likelihood. For dorsals, the opposite is true; OCP-DORS is significant, but OCP-DORS-SON is not.

Results confirm that transvocalic consonant OCP effects are active in Samoan root phonotactics. Consistent with the literature on OCP-place, I also find effects of subsidiary

features. Notably, existing work on OCP-place disagrees on how much the effect of subsidiary features should be allowed to vary across places of articulation. Coetzee & Pater (2008b) allow for the weights of subsidiary features to vary across place, while Wilson & Obdeyn (2009) and Frisch et al. (2004) both argue for more restrictive implementations.

A comprehensive comparison of these different theories is beyond the scope of the current study. However, it should be noted that the Samoan results appear to support Coetzee and Pater's less restrictive approach, since the effect of SONORANT is different across places of articulations and much stronger for coronals.

4.2.3   OCP effects in Proto-Oceanic and Proto-Polynesian

Based on a crosslinguistic study of six languages, Krupa (1971) proposes that OCP-place is a general property of Polynesian languages. Other work, such as Mester (1986) and Coetzee & Pater (2008b), also suggests that OCP-place effects may be a general property of Austronesian languages. Samoan closely resembles Proto-Polynesian and has not undergone place-related sound changes with the exception of *k>ʔ. This means that if OCP-place restrictions are active in modern Samoan, they were likely present (and therefore able to affect reanalysis) at an earlier stage of the language.

In fact, a survey of Proto-Polynesian (PPn) suggests that the same phonotactic restrictions present in Samoan already existed in an earlier stage of the language. To test for OCP-place effects in PPn, a corpus of PPn protoforms was collected from POLLEX (Greenhill & Clark 2011). This data was filtered to remove reduplicated forms and compounds. Words were also stripped of common affixes (e.g. *faa-, *faka 'CAUSATIVE, *fe- 'RECIPROCAL', *ma- 'STATIVE', *-Caŋa, *ŋa 'NOMINALIZER'). The resulting corpus of 1645 PPn protoforms was used to produce Fig. 4.5, which shows consonant-consonant co-occurrences in PPn. For comparability with the Samoan data, consonants are grouped by their reflex in Samoan, rather than the actual PPn reconstructions.

The boxes frame regions where OCP-place effects were found for Samoan, and there-
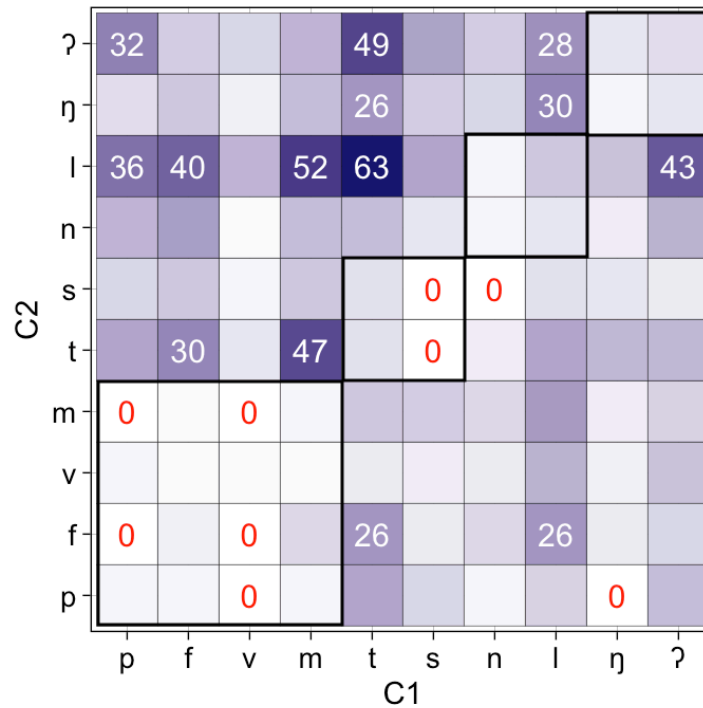
Figure 4.5: Consonant-consonant co-occurrences in PPn

fore where C1-C2 pairs are expected to be underrepresented. We can see that in general, the PPn distributions match the Samoan distribution.

This is confirmed using a MaxEnt model with the same structure as in §4.2.2 above. Again, OCP-place constraints were tested for significance using the Likelihood Ratio Test (Hayes et al. 2012). The results, given in Table 4.4, are consistent with the findings for the modern Samoan data. In particular, OCP-LAB, OCP-COR-SON, and OCP-BACK tested as significant, and these were the same three constraints found to be significant for Samoan.

Constraint	w	ΔL	p
OCP-place	0.09	0.005	n.s. (0.92)
OCP-LAB	1.77	33.83	1.95×10^{-16}***
OCP-LAB-SON	0.12	0.03	n.s. (0.81)
OCP-COR	0.06	0.99	n.s (0.16)
OCP-COR-SON	1.29	30.15	8.12×10^{-15}***
OCP-BACK	0.75	3.97	0.005**
OCP-BACK-SON	0.41	0.01	n.s. (0.36)

Table 4.4: OCP constraint weights learned by the phonotactic model for PPn

112

Examination of POc, which represents an even earlier stage of the language, suggests that there might be a mismatch between the OCP-place effects found in Polynesian vs. in the Oceanic languages as a whole. This becomes important in the following section, as it results in a mismatch between stem phonotactics and cross-morpheme phonotactics in Polynesian languages such as Samoan; these mismatches become relevant when we look at patterns of thematic consonant reanalysis later on.

Fig. 4.6 shows consonant-consonant co-occurrences in POc, using data from the ACD (Blust & Trussel 2010). The right-hand figure shows the subset of all C1-C2 pairs after excluding word-final consonants; Proto-Polynesian (PPn) regularly lost all final consonants, so a C1-C2 set which excludes final consonants is likely a closer reflection of PPn stem phonotactics.
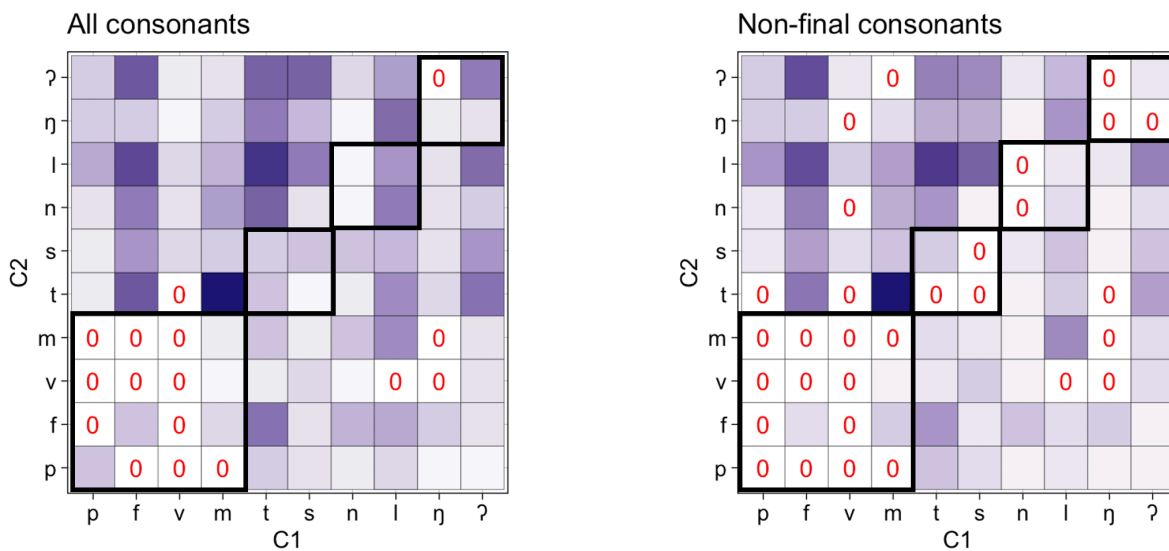
Figure 4.6: Consonant-consonant co-occurrences in POc

In the right-hand figure (with final consonants removed), OCP-place effects appear to be present. In the left-hand figure, which includes final consonants, there appears to be an effect of OCP-LABIAL, but not of the other OCP-place constraints. Taken together, these results suggest that in POc, OCP-place effects were present in earlier parts of the stem, but not when stem-final consonants were accounted for. The subsequent deletion of final consonants in PPn therefore resulted in stronger OCP-place effects.

Thematic consonants originate from POc final consonants. As such, it is conceivable that in Samoan (and other Polynesian languages), there would at some point have been a mismatch in the strength of OCP-place effects within stems and across morpheme boundaries (i.e. in suffixed forms where thematic consonants surface). In fact, this turns out to be the case. In the following section, I show that in an earlier stage of Samoan, there were thematic consonant alternations which violated OCP-place constraints. Over time, these forms were reanalyzed at a higher rate than thematic consonant alternations which did not violate OCP-place, demonstrating the effects of a markedness bias.

4.3 Reanalysis in Samoan

In this section, I present evidence for reanalysis in Samoan. Two types of data are compared: POc protoforms (representing thematic consonants *before* reanalysis) and modern Samoan stem-suffix pairs (representing the state of Samoan after reanalysis had occurred). §4.3.1 first presents the distribution of final stops in POc and predicted directions of reanalysis.

POc reconstructions are taken from the Austronesian Comparative Dictionary (ACD Blust & Trussel 2010). Protoforms were excluded if they had fewer than six cognates within Oceanic (i.e., not counting cognates from other Austronesian sub-familiies), resulting in a set of 1023 protoforms.

I also tested a separate, more restricted list of 279 protoforms, which were filtered to include only forms reflected in Samoan. The rest of this chapter will report results from the larger corpus, but note that similar results were found using this smaller set of protoforms.

Modern Samoan forms are taken from the Milner (1966) dictionary and supplemented with forms from Pratt (1862/1893). I focus on stem-ergative pairs, since of all the suffixes that trigger thematic consonant alternations, the ergative is the most productive and has more available forms. The resulting wordlist has 593 stem-suffix pairs.

4.3.1 Distribution of final consonants in POc

Since thematic consonants developed from POc final consonants, looking at the distribution of final consonants in POc can give us insight into the expected distribution of ergative allomorphs, *before* reanalysis had occurred. Table 4.5 shows the distribution of final consonants in POc and the expected ergative allomorph given this final consonant.

Around 68% of stems were either historically vowel-final, or ended in a consonant that was uniformly deleted by regular sound change (and therefore should not have thematic consonant alternations). More concretely, around 68% of words are expected to take one of the vowel-initial ergative allomorphs (/-a/ or /-ina/). This means that in a frequency-matching model, we should expect reanalysis towards /-a/ and /-ina/.

POC	Allomorph	Count	Total
*vowel,*ʔ,*R,*y	(in)a	189	0.68
*p	fia	9	0.03
*d,*l,*r	lia	8	0.03
*m	mia	6	0.02
*n,*ɲ	na, ina	24	0.09
*ŋ	ŋia	12	0.04
*s	sia	7	0.03
*t	tia	13	0.05
*k	ʔia	11	0.04

Table 4.5: Distribution of final consonants in POc

Fig. 4.7 shows the expected distribution of ergative allomorphs by the immediately preceding consonant, based on data from POc. For ease of reading, vowel-final protoforms (corresponding to suffixes /a/ and /ina) are ommitted. First, we can observe a strong effect of OCP-place for labials; where the preceding consonant is one of /p, f, v, m/, the expected ergative allomorph is never /fia/ or /mia/. In other words, forms like [la**pa**-**fi**a] or [ta**ma**-**mi**a] are never observed. For coronals and dorsals, the OCP-place effect is weaker. Most strikingly, forms of the type [i**la**-**na**], where preceding /l/ is followed by /-na/, are relatively frequent (n = 11).

If reanalysis is predictable from statistical distributions within a paradigm, we might

115
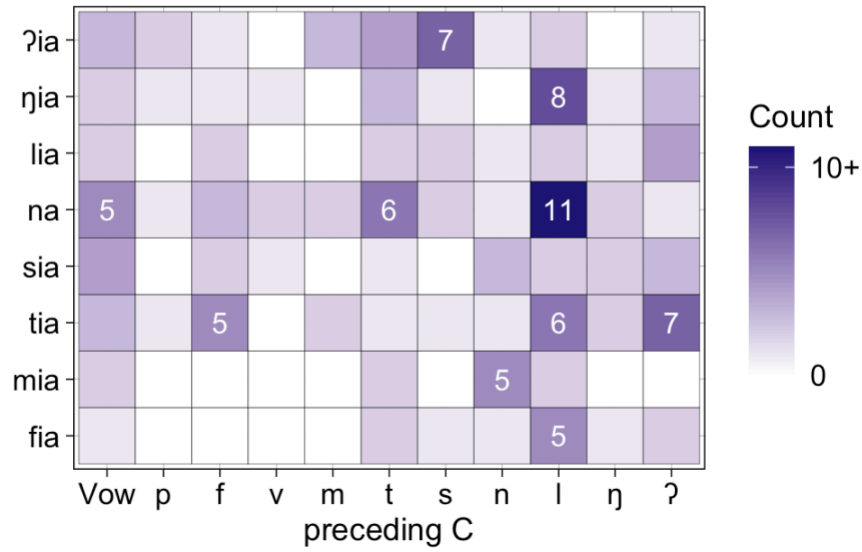
Figure 4.7: Expected distribution of ergative allomorphs by identity of preceding consonant (POc)

expect reanalysis to occur in a way that avoids labial-labial sequences, but not sequences of coronal sonorants.

4.3.2 Comparing POc and Samoan

In this section, I compare the distributional patterns of the thematic consonant in POc against modern Samoan data. This comparison provides indirect insight into the direction of reanalysis, where mismatches between POc and Samoan suggest that reanalysis has occurred in a way that is not fully predictable from frequency-matching models. The following section will then provide a form-by-form comparison of reanalyses that have actually been observed.

First, Fig. 4.8 compares the overall distribution of allomorphs in POc and Samoan. In general, the two are closely matched, as predicted by the frequency-matching approach to reanalysis. Note that the modern Samoan data may under-estimate the proportion of stems which take /-a/ and /-ina/, since loanwords and other innovative forms that are omitted from the data will generally take /-ina/. Additionally, Pratt (1862/1893) does

116

not list passive forms if they end in /-ina/. Even if this is the case, reanalysis would still be towards the majority variants, in line with frequency-based models.
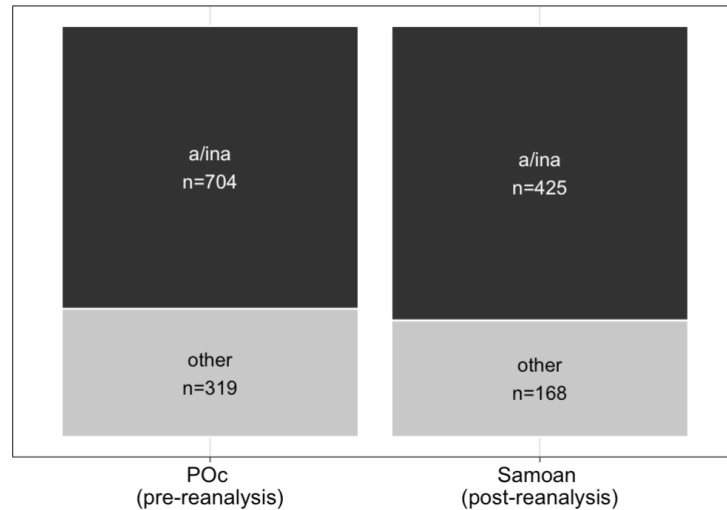
Figure 4.8: Distribution of ergative allomorphs before and after reanalysis

Fig. 4.9 compares the distribution of ergative allomorphs in POc and Samoan by identity of the preceding segment; forms which take /-a/ and /-ina/ are omitted. Some distributional patterns present in POc are carried over to Samoan. For example, the effect of OCP-labial was exceptionless in POc, and this is carried over to modern Samoan. On the other hand, stems of the type [ilo-na] (where the suffix allomorph is [na], and the preceding consonant of the stem is [l]) are never attested in Samoan despite their relatively high frequency in the POc data.

I test whether [ilo-na] type stems are underrepresented in Samoan, given the POc distribution, using a Monte Carlo test of significance. First, for every POc protoform, I extracted the preceding consonant (C_{prev}) and final consonant (i.e. thematic consonant, C_{theme}). To limit the number of comparisons, consonants were then collapsed into natural classes based on combinations of place ([LABIAL, CORONAL, DORSAL]) and manner (sonorant vs. obstruent). For example, [COR,SON]-[DORS,SON] covers protoforms like *buliŋ (C_{prev} = [l], C_{theme} = [ŋ]) and *baniŋ 'bait' (C_{prev} = [n], C_{theme} = [ŋ]).

I then randomly recombined the extracted consonants to make new C_{prev}-C_{theme} pairs.
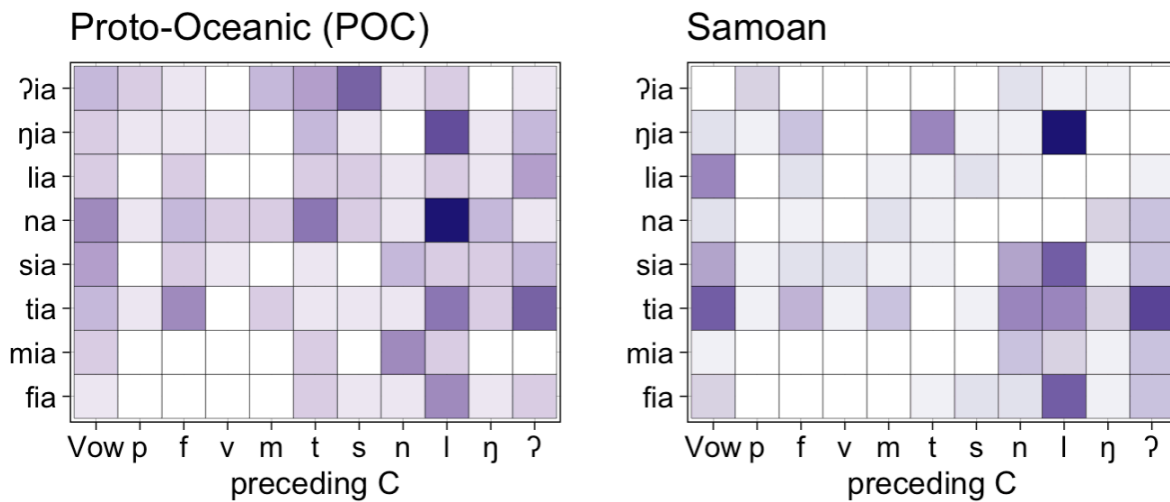
Figure 4.9: Distribution of allomorphs by preceding consonant in POc vs. Samoan

This process was repeated 10,000 times to produce the expected chance-level distribution of each C_{prev}-C_{theme} pair in POc. The observed count of each C_{prev}-C_{theme} pair in modern Samoan is then compared against this distribution.

The POc corpus has 1023 forms and the Samoan corpus has 593 stem-ergative pairs. Because the two sets of data differ in size, I scaled the Samoan C_{prev}-C_{theme} counts by randomly sampling 1023 forms 10,000 times and taking the average count from all trials.

Fig. 4.10 demonstrates how interpret the Monte Carlo results. The interval shows the 95% confidence interval for [COR, SON]-[DORS, SON] pairs derived from the Monte Carlo test. It represents the expected distribution of data in POc, pre-reanalysis. The dot represents the actual attested counts in Samoan of stems where the ergative allomorph /-ŋia/ is preceded by a coronal sonorant. In this specific example, the Samoan count is larger than the 95% confidence interval, meaning that stems of the type [ina-ŋia] and [ila-ŋia] are over-attested in Samoan, given the historical POc distribution.

Fig. 4.11 visualizes the Monte Carlo results for the subset of C_{prev}-C_{theme} pairs where at least one segment is a coronal sonorant. This figure essentially compares the expected distribution given POc against the observed counts in Samoan for these C_{prev}-C_{theme} pairs. Most C_{prev}-C_{theme} pairs are either over-attested or within the expected range given chance.

118
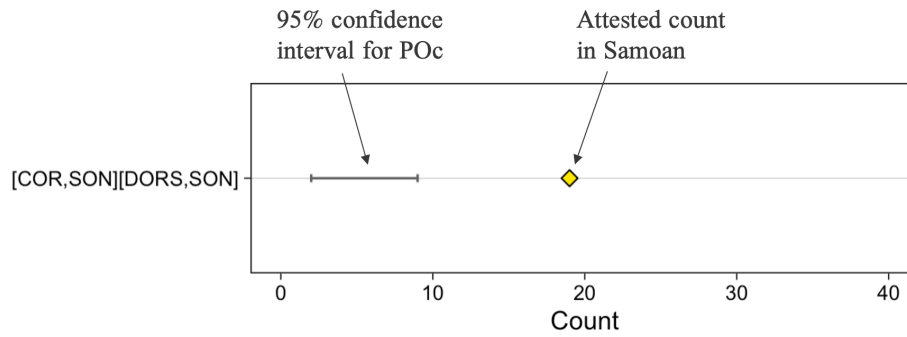
Figure 4.10: Chance-level distribution of C_{prev}-C_{theme} pairs vs. observed count in Samoan
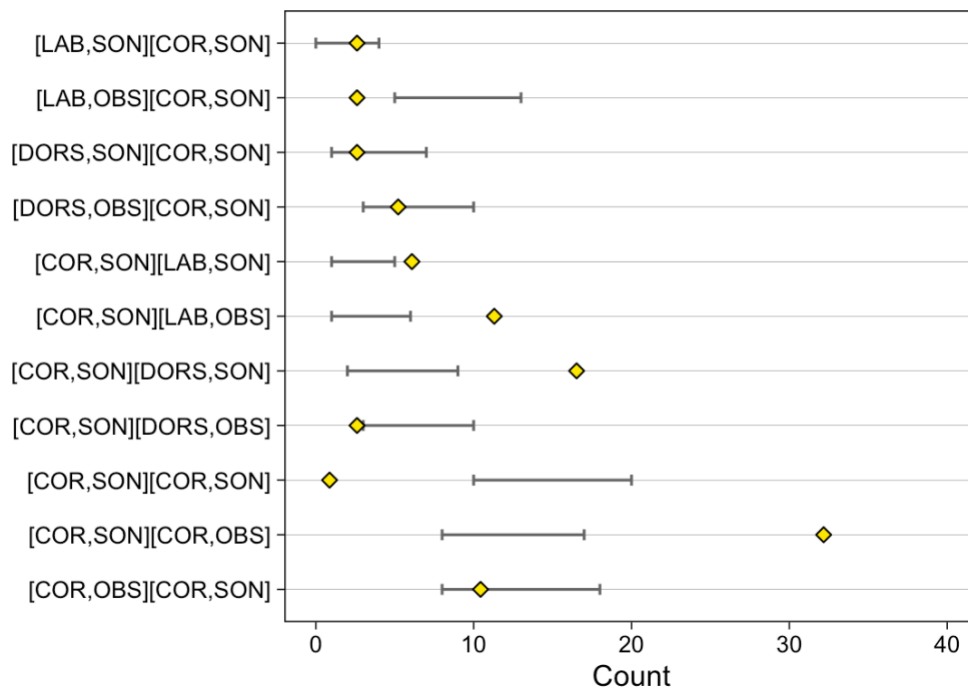
Figure 4.11: Chance-level distribution of C_{prev}-C_{theme} pairs vs. observed count in Samoan

For each POc form, the most frequent reflex is shown in bold

[LAB,OBS][COR,SON] pairs (e.g. [ipa-lia], [ifa-na]) are slightly underattested. Most strikingly, however, [COR,SON][COR,SON] pairs are highly under-attested.

Overall, comparison of POc and Samoan suggests that reanalysis is generally in the direction of the statistically most common variants (/a/, /-ina/). However, I propose that reanalysis is additionally sensitive to OCP-place effects. In particular, stems like [ino-lia]

119

and [ilo-na] are more likely to be reanalyzed because violate the markedness constraint OCP-COR-SON.

4.3.3 Direct evidence of reanalyses

In this section, I consider the subset of stem-ergative pairs which have known POc proto-forms (n = 147). These forms provide direct evidence for reanalyses that have occurred. Table 4.6 summarizes the proportion of forms that have undergone reanalysis based on what allomorph they would have taken historically. Fig. 4.12 visualizes this same data. Overall, results are consistent with the conclusions of the previous section.

In particular, reanalysis mostly is towards /-a/ and /-ina/ (labeled here as /-(in)a/). First, Samoan stems that are expected to take /(in)a/ have undergone the least amount of reanalysis; just 8% of these have been reanalyzed. In contrast, for stems that histori-cally took a /-Cia/ allomorph, there has been more extensive reanalysis. Finally, where reanalysis has occurred, it is most often towards /-(in)a/ rather than another allomorph.

POc	Samoan	n	%	POc	Samoan	n	%
(in)a (n=72)	(in)a	66	0.92	na (n=9)	na	4	0.44
	other	6	0.08		(in)a	3	0.34
					other	2	0.22
fia (n=8)	fia	6	0.74	sia (n=10)	sia	6	0.6
	(in)a	1	0.13		(in)a	3	0.3
	other	1	0.13		other	1	0.1
ŋia (n=9)	ŋia	3	0.33	tia (n=19)	tia	9	0.47
	(in)a	5	0.56		(in)a	9	0.47
	other	1	0.11		other	1	0.06
lia (n=9)	lia	2	0.22	ʔia (n=6)	ʔia	2	0.33
	(in)a	5	0.56		(in)a	4	0.67
	other	2	0.22		other	0	0
mia (n=4)	mia	3	0.75				
	(in)a	1	0.25				
	other	0	0				

Table 4.6: Summary of reanalyses (POc protoforms vs. Samoan reflexes)

Table 4.7 breaks down reanalysis involving /-lia/ and /-na/ forms by whether the
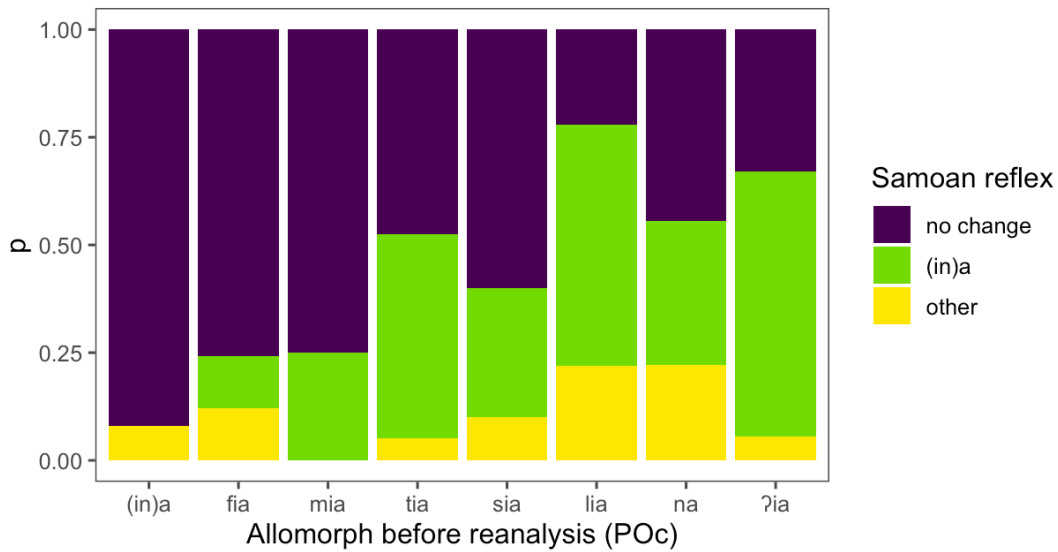
Figure 4.12: Summary of reanalyses (POc protoforms vs. Samoan reflexes)

stem has an immediately preceding /l/ or /n/. Again, where the expected allomorph given POc mismatches the modern Samoan one, reanalysis has occurred. Overall, results are consistent with the previous section. In all cases where the stem had a preceding /n/ or /l/, reanalysis occurred; this is true for both stems that were expected to take /-lia/ and ones expected to take /-na/. I argue that this is because words of the type /i**n**a-**l**ia/ and /i**l**a-**n**a/ violate OCP-COR-SON.

POc	has /n,l/	Samoan	N	Example
lia	yes	lia	0	[ana-lia] (<*anal)[a]
	yes	other	4	[fono-tia] (<*ponor, cf. *[fono-lia]) 'hold meeting'
	no	lia	2	[tautau-lia] (<*saur) 'hang up'
	no	other	3	[fata-ina] (*pantar, cf. *[fata-lia]) 'carry in bag'
na	yes	na	0	[ali-na](<*anlin)[a]
	yes	other	4	[talu-ina] (<*talun, cf. *[talu-na]) 'weed growth'
	no	na	4	[aŋi-na] (<*haŋin) 'to blow'
	no	other	1	[leːnifo-a] (<*nipon, cf. *[leːnifo-na]) 'tooth'

Table 4.7: Reanalyses of /lia/ and /na/

[a]Hypothetical example (no real examples fall into this category)

4.3.4 Interim summary

Overall, the picture that emerges is similar to what was found for Malagasy, where a mismatch between stem phonotactics and morphophonological alternations was gradually removed over time. Modern Samoan has strong OCP-place effects, specifically of OCP-LABIAL, OCP-COR-SON, and OCP-BACK. In POc, effects of OCP-COR-SON, and OCP-BACK only emerged after omitting final consonants, suggesting that these constraints were not active in final consonants.

It is plausible that at some point in the history of Samoan following the loss of final consonants, OCP-place effects were active in stems, but there was a mismatch between stem phonotactics and thematic consonant alternations. Over time, markedness-sensitive reanalysis removed the OCP-violating alternations. In the following section, I show that markedness-biased models, specifically ones that incorporate OCP-place effects, outperform purely frequency-matching ones.

4.4 Modeling Samoan reanalysis

In this section, I test whether reanalysis of Samoan thematic consonants can be modeled as the combined effects of frequency and markedness, using the model outlined in Chapter 2. §4.4.1 and §4.4.2 describe aspects of the model implementation that are specific to Samoan and §4.4.3 presents model results.

In addition, §4.4.2 will introduce several different phonotactic models, which will then be compared on how well they predict the direction of reanalysis in Samoan. Malagasy (discussed in Chapter 3) had a relatively simple alternation pattern involving just two alternants for each environment. As a result, all phonotactic models performed equally well. Samoan, on the other hand, is a more complicated case where speakers, when given a stem, are forced to pick between many competing allomorphs. This allows us to test for more nuanced effects in how stem phonotactics influences reanalysis.

4.4.1 Choice of URs and inputs

As famously pointed out by Hale (1968, 1973), the Polynesian thematic consonant has two analyses: under the so-called phonological analysis, the thematic consonant belongs to the stem UR and is deleted in unsuffixed forms by a regular phonological rule of final consonant deletion. As shown in the third column of (43), this approach allows the ergative suffix to have one predictable allomorph. Under the second 'morphological' analysis, suffixes have multiple suppletive allomorphs and roots are marked for which ones they take. This approach, shown in the fourth column of (43), makes morphophonology more complex. On the other hand, stem URs are closer to their surface forms.

(43) *URs under phonological vs. morphological analysis*

		UR	
SRs	gloss	phonological	morphological
[inu]∼[inumia]	'to drink'	/inum/+/ia/	/inu/+/mia/
[ita]∼[itaŋia]	'to be angry'	/itaŋ/+/ia/	/ita/ + /ŋia/
[piʔi]∼[piʔitia]	'to cling'	/piʔit/+/ia/	/piʔi/ + /tia/

Hale (1968, 1973) looked specifically at Māori and argued in support of the morphological analysis. Since then, there has been extensive debate about the status of thematic consonants, with some in favor of the phonological analysis (e.g. Sanders 1990; de Lacy 2002, 2003) and others in support of the morphological analysis (e.g. Blevins 1994; Lichtenberk 2001).

For Samoan (and Māori in the following chapter), I adopt the morphological analysis. The reasons, many of which follow from Hale's analysis of Māori, are as follows: first, for the same stem, different thematic consonants may surface in different suffixal contexts. Some examples are given in (44) with the thematic consonant shown in boldface (and ∅ where no thematic consonant surfaces). This suggests that the thematic consonant underlyingly belongs to the suffix, rather than the stem.

123

(44) *Variation in thematic consonant across suffixal contexts*

STEM	SUFFIXED FORMS	GLOSS
alofa	alofaŋia, alofaʔaŋa	'love, affection'
eʔe	eʔetia, eqenaʔi	'be propped up/raised'
au	aulia, au∅aʔi	'flow on, continue
tae	taːe∅a, taenaʔi	'gather'
sua	sua∅ina, suataŋa	'lever (up)'

Additionally, some stems can take multiple allomorphs for the same suffixal context. In some cases, the meaning of the derived form will differ depending on the allomorph, as demonstrated in (45) for the stem [tuʔu].

(45) *[tuʔu] with different ergative suffix allomorphs*

tuʔu-ina 'give, grant'
tuʔu-a   'left, depart from, refuse'
tuːʔu-a   'dismiss'
tuʔu-na   'leave behind'

An example of a model input and its corresponding candidates is given in (46). Essentially, the input is unsuffixed stems, while candidates take different suffix allomorphs. Because the thematic consonants are analyzed as belonging to the suffix, stems have a transparent UR that matches the SR.

(46) *Example of model input and candidates*

INPUT	CANDIDATES
[inu]∼/inu/-ERG	inu**(in)a**
	inu**fia**
	inu**mia**
	inu**tia**
	inu**sia**
	inu**na**
	inu**lia**
	inu**ŋia**
	inu**ʔia**

Model inputs are 500 stems whose distribution reflect that of the POc protoforms. In selecting the model inputs, I also make several simplifying assumptions. For the suffixed form candidates, /-ina/ and /a/ are combined, since their relative historical distribution is unclear. Inputs are also pooled by the identity of the preceding consonant (/p,f,v,m,t,s,n,l,ŋ,ʔ/ or 'none'). I do not consider conditioning effects of stem shape or final vowel. Therefore, an input like /ino/ represents all stems where the preceding consonant is /n/.

Note that although I adopt the morphological approach, both the phonological and morphological approaches can be used to model reanalysis. Under the morphological approach, the learner's goal is to pick between possible allomorphs. The choice between different allomorphs can be enforced using violable morpheme exponence constraints of the form 'ERG = /tia/', which demand a particular exponent for a particular morphological category (Russell 1995b; Kager 1996).

This is shown in the tableau in (47), which for illustrative purposes uses hand-fitted constraints and a simplified candidate set. In this tableau, ERG = /(in)a/ has a relatively high weight, reflecting its status as the most frequent (default) allomorph. Consequently, candidate (a) has the highest predicted probability. Exponence constraints can also interact with markedness constraints. In this example, OCP-COR-SON penalizes candidate (b), causing it to be assigned the lowest predicted probability.

(47)   *Tableau: morpheme exponence constraints*

	ERG = /(in)a/	ERG = /na/	ERG = /tia/	OCP-COR-SON	$\mathcal{H}$	P
	3	0.5	0.5	1		
/pili-ERG/						
a. pili-a		1	1		1.00	0.90
b. pili-na	1		1	1	4.50	0.03
c. pili-tia	1	1			3.50	0.07

In contrast, under the phonological approach, the learner's goal is to decide what a stem's UR should be, given an incomplete paradigm. In other words, for a stem like [pili] 'to be stuck', if the learner has never heard a suffixed form, they must decide whether the stem UR is /pili/, /pilin/, /pilit/, etc. Although I do not adopt the phonological analysis, this process can be modeled using UR inference constraints (outlined in Chapter 2).

Tableau (48) illustrates what UR inference would look like for [pili]. In this model, the input is not phonological material, but rather is something like word meaning, encoded as morphosyntactic features and associated surface forms. The candidates are UR-SR pairs, and constraints enforce different SR-UR mappings. For example, [V#] requires that a final vowel be underlyingly /V/, and therefore penalizes candidates (b) and (c). Since the candidates are UR-SR pairs, they are also sensitive to surface markedness constraints. For example, candidate (b) violates OCP-COR-SON and is therefore assigned a lower predicted probability than the other candidates.

(48) *Tableau: UR inference constraints*

	[V#] = /V/	[V#] = /Vl/	[V#] = /Vt/	OCP-COR-SON	$\mathcal{H}$	P
	3	0.5	0.5	1		
|STUCK, [pili]| + |ERG,[ia],[a]|						
a. /pili-ia/ [pilia]		1	1		1	0.90
b. /pilil-ia/ [pililia]	1		1	1	4.5	0.03
c. /pilit-ia/ [pilitia]	1	1			3.5	0.07

4.4.2   Implementing a phonotactic markedness bias

To implement markedness bias, I follow the steps schematized in Fig. 4.13. First, phonotactic grammars are trained on monomorphemic roots using the UCLA Phonotactic Learner (UCLAPL; Hayes & Wilson 2008). The UCLAPL is itself based in MaxEnt; it learns weights for phonotactic constraints and can be used to assign Harmony scores to words (where the higher the Harmony, the more phonotactically marked a word is). Using the grammar learned by the UCLAPL, I assign harmony scores to the candidate suffixed forms of the model of reanalysis. These harmony scores then become the constraint violations for a constraint USEPHONOTACTICS; this is the constraint that is given a bias towards high weight.

This single USEPHONOTACTICS constraint essentially aggregates all the constraints from the phonotactic grammar, while fixing their relative weights to be the same as they were in the phonotactic grammar. Using this method, markedness effects can be derived directly from root phonotactics, without the need to stipulate specific constraints.

The input to the phonotactic model is a corpus of 1645 PPn protoforms taken from POLLEX (Greenhill & Clark 2011); this is the same corpus used earlier in §4.2. As de-

1. Root phonotactics

[inu]
[tapa]
[manu]
[tai]
...
→ Phonotactic grammar

2. Morphophonology

Candidates:
[ilotia]
[iloŋia]
[Ilona]
[ilolia]
...

→ Phonotactic grammar →

Penalty (H)	
[ilotia]	0.23
[iloŋia]	0
[Ilona]	1.13
[ilolia]	1.13
...	

→

	USEPHON	C1	C2...
[ilotia]	0.23	1	0
[iloŋia]	0	0	1
[ilona]	1.13	0	0
[ilolia]	1.13	0	0
...			
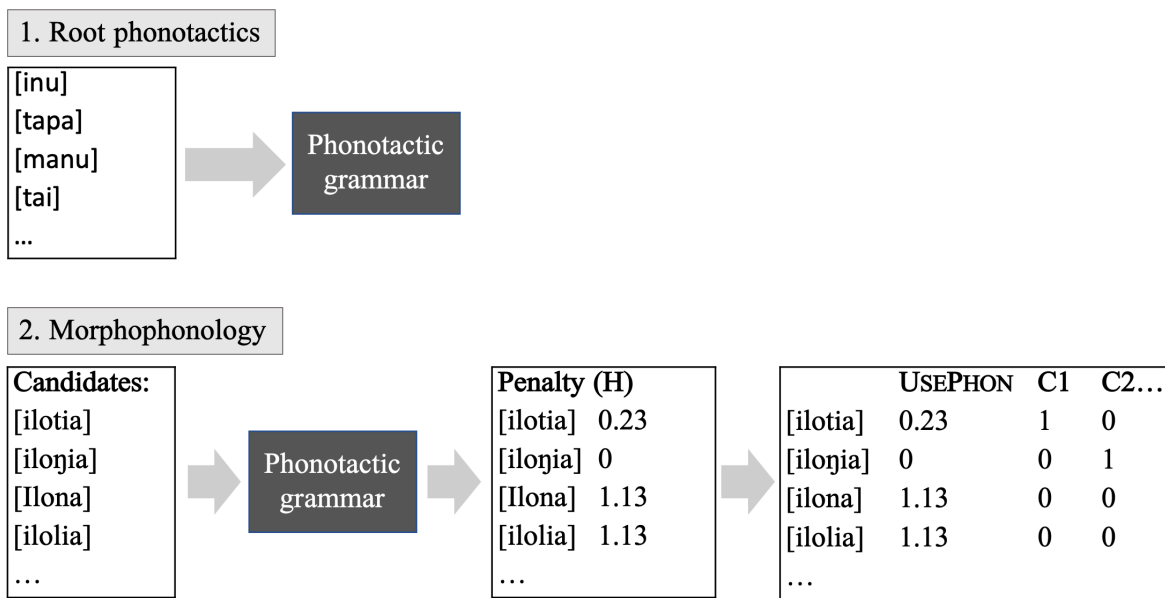
Figure 4.13: Incorporating phonotactic markedness into morphophonological grammar

scribed there, this corpus had polymorphemic items removed, and was modified to reflect the regular sound changes that have happened between PPn and Samoan. The model is also given a fairly standard feature set, with one major difference: each diphthong is represented as a single phoneme (with a main vowel and offglide), and diphthongs are distinguished from other vowels using the constraints [high_glide] and [back_glide], which specify the [back] and [high] features of the diphthong's offglide. These constraints, while non-standard, are enough to distinguish between all vowel categories. There is also a feature [long], which separates short vowels ([-long]) from long vowels and diphthongs ([+long]).

The UCLAPL can discover its own constraints using a set of search heuristics, or it can learn weights on a set of pre-specified constraints. I trained the following four phonotactic models using the same corpus of 1645 roots used in §4.2:

1. NATURAL CLASS MODEL: This model was limited to learning 50 constraints and given no prespecified constraints. In addition, the model was given a consonant projection (which includes all [-syllabic] segments).

2. OCP MODEL: This model was given a set of prespecified constraints, consisting of all possible combinations of OCP-place constraints (OCP-LABIAL, OCP-CORONAL, and OCP-BACK) with the subsidiary features [sonorant], [voice], and [continuant]. Where constraints were redundant (i.e. targeted the exact same class of segments), one of the constraints was removed.

3. OCP + HIATUS MODEL: This model was given prespecified constraints, consisting of the same OCP constraints used in the above model, along with five hiatus-related constraints, given in (49). These constraints are discussed in more detail in Chapter 5, but are included for two purposes: first, they are used to account for possible markedness effects incurred by the vowel-initial suffixes (/-ia/, /-a/, and /-ina/). Additionally, as a preview of the results in Chapter 6, vowel hiatus effects are found to be active in the reanalysis of Māori thematic consonants. Since Māori and Samoan have similar origins (both belonging to the Polynesian language family), the same effects are potentially active in Samoan.

4. BIGRAM MODEL: This model was given a set of prespecified constraints, consisting of all possible C1-C2 combinations.

(49) *Hiatus constraints*

Constraint	Example penalized outputs
*[+long]	h**aː**, t**ai**
*[+syllabic][+syllabic]	t**ia**, t**ie**
*[+word_boundary][+syllabic]	**a**ʔe, **e**va, **e**a
*[+long][+syllabic]	t**iːa**, t**auaː**
*[+syllabic][+long]	t**iai**, t**eːaː**

The four models are summarized in (50) by the number of parameters they each have, which is generally defined as the number of constraints. The OCP model has 19 constraints, but has one additional parameter to account for the restriction that constraints

129

must target C1…C2 pairs, where C1 and C2 share the same feature specifications. Notably, the OCP model has the fewest constraints and is also the most restrictive in terms of how constraints can be defined.

(50) *Summary: phonotactic grammar parameters*

Model	Parameters
NATURAL CLASS	50
OCP	20
OCP + HIATUS	26
BIGRAM	100

These models are compared to help us gain insight into how speakers' phonotactic knowledge can affect reanalysis. The NATURAL CLASS, OCP, and OCP + HIATUS models allow generalization to natural classes, while the BIGRAM model doesn't. If the BIGRAM model outperforms the other models, this suggests that speakers are simply learning C1-C2 probabilities and applying this to resolve ambiguities in an alternation pattern.

On the other hand, if the NATURAL CLASS and OCP (+ HIATUS) models perform better, this suggests that speakers prefer to generalize patterns to natural classes. The OCP and OCP + HIATUS models are additionally more restrictive, in that they only allows for typologically-motivated constraints, rather than potentially arbitrary constraints learned over any natural class. If the OCP (+ HIATUS) models outperform the other models, this suggests that speakers do not pick up on any phonotactic regularity in the lexicon, but prefer to learn more well-motivated constraints.

4.4.3 Model specifications and results

All model results were averaged over 30 runs, and each model was run for 20 iterations. σ was set to 1 for all constraints.

In the markedness-biased models, the USEPHONOTACTICS constraint is given a μ value of 3; $\mu = 0$ for all other constraints. These models are compared against a BASELINE

model where all constraints are given a μ value of 0. For completeness, three baseline models were tested, each with an USEPHONOTACTICS constraint derived from one of the three phonotactic models. Since they all behaved very similarly (±1 in log-likelihood), the following model results show just the baseline model with an USEPHONOTACTICS constraint derived from the NATURAL CLASS phonotactic grammar.

Table 4.8 compares the log-likelihood of each model. The rightmost column (ΔL) shows the change in log-likelihood of each model compared to the baseline. Overall, all four markedness-biased models outperform the BASELINE model.

	L	ΔL
BASELINE	-2448.81	–
NATURAL CLASS	-2416.27	32.54
OCP	**-2385.00**	**63.81**
OCP + Hiatus	**-2383.20**	**65.61**
BIGRAM	-2438.39	10.42

Table 4.8: Model results: log likelihood

Of the markedness-biased models, the BIGRAM model performs the worst, while the two models with OCP constraints do the best. This suggests that models which generalize to natural classes are better predictors of learner behavior. The OCP + HIATUS model does marginally better than the OCP model, but also has more constraints; the difference between the two models also comes out as non-significant in a Likelihood Ratio Test.

Notably, the NATURAL CLASS grammar is able to generalize to natural classes, but still does not perform as well as the OCP grammar. A closer inspection of the two (OCP vs. NATURAL CLASS) suggests that the NATURAL CLASS grammar learns constraints that are still not sufficiently general, especially for the coronal sonorants.

Based on the POc data, stems of the type [**ino-n**a], [**ino-l**ia], and [**ilo-l**ia] are expected to be infrequent, but [**ilo-n**a] stems were relatively frequent. Because of this, the NATURAL CLASS grammar does not learn a general OCP-coronal sonorant constraint, but instead learns the three separate constraints given in (51). The constraint *[l...n] is assigned a relatively lower weight, so the model does not penalize [ilo-na] type words as heavily

and under-predicts the rate at which they are reanalyzed. In contrast, the OCP model is forced to learn a more general OCP-COR-SON constraint, and therefore assigns a higher penalty to [ilo-na] type words.

(51)   *Constraints on coronal sonorant C1-C2 pairs in the NATURAL CLASS grammar*

CONSTRAINT	W	CANDIDATES PENALIZED
*[n...{l,s}]	1.26	ino-lia, ino-sia
*[{l,n}...l]	0.93	ino-lia, ilo-lia
*[l...n]	0.78	ilo-na

Fig. 4.14 compares predictions of the BASELINE and OCP models for stems with a preceding [l] (i.e. inputs of the type [ilo]). For ease of interpretation, only a subset of candidates are included. For [ilo]-type stems, the biggest different between POc and Samoan is that Samoan has a much lower proportion of the candidate [ilo-na]. The baseline model is unable to predict this, while the OCP model can (since again [ilo-na] violates OCP-COR-SON).
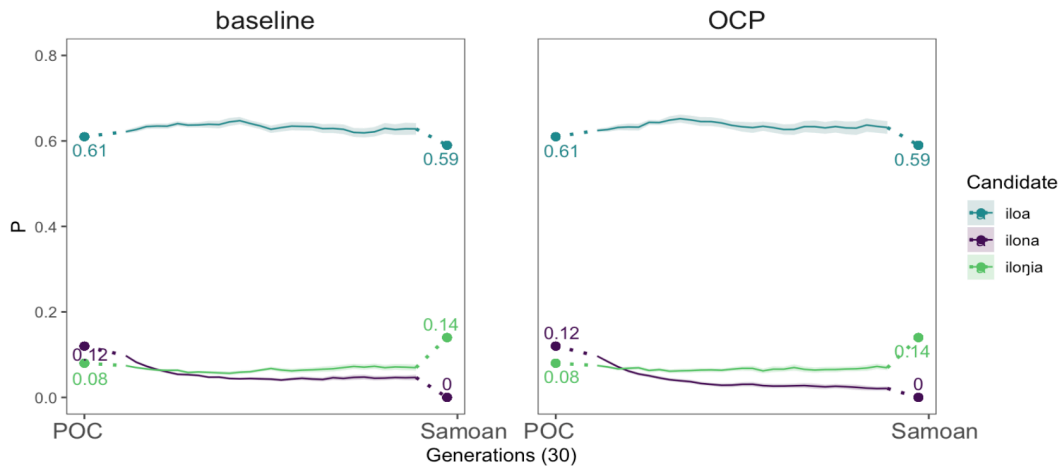
Figure 4.14: Model predictions in stems with a preceding /l/

4.5   Chapter conclusion

Overall, results in this section complement the findings for Malagasy, where a model of reanalysis that incorporates a phonotactic markedness bias outperform purely frequency-matching models.

In this section, I also compared between three phonotactic models. All three models were based in the UCLAPL and designed to match statistical patterns in Samoan roots. However, the models differ in terms of the types of constraints that can be learned. The OCP model, which allows for generalization to natural classes but is also very restrictive about the possible range of constraints, actually performs the best. This suggests that while speakers draw on phonotactic knowledge when resolving ambiguities in paradigms, they do not pick up on all patterns, but instead are biased to pick up on more general patterns rooted in markedness motivations.

CHAPTER 5

Case study 3: Māori thematic consonants

Māori is a Polynesian language of the Nuclear Polynesian family, spoken in mainland New Zealand. Like Samoan, Māori has thematic consonant alternations. In Māori, thematic consonant alternations surface mostly in the passive suffix; like the Samoan ergative discussed in Chapter 4, the Māori passive also descends from the PPn CIA suffix. Other suffixes which historically carried the thematic consonant are either non-productive or only marginally productive.

In Samoan, the default allomorphs of the CIA suffix are the vowel-initial /-a/ and /-ina/. In Māori, however, the most common passive allomorphs are /-a/ and /-tia/. The shift towards /-tia/ as a productive allomorph is surprising because based on the historical distribution of final segments in POc, vowel-initial allomorphs should be the most frequent. In this chapter, I propose that reanalysis towards /-tia/ as a default alternant was motivated by markedness considerations. In particular, /-tia/ is adopted to avoid violation of two markedness constraints, *LONGNUCLEUS and NOONSET, which interact to penalize hiatus involving heavy syllables.

Section 5.1 introduces the Māori thematic consonant alternation pattern, with a focus on where it differs from the Samoan pattern discussed in Chapter 4. Section 5.2 presents evidence for hiatus avoidance in Māori and PPn stem phonotactics. Following this, Section 5.3 presents data on reanalysis in Māori and evidence for the effect of hiatus-related constraints on reanalysis. Section 5.4 presents modeling results which show that reanalysis towards /-tia/ can be predicted by a markedness bias rooted in stem phonotactics. Finally, §5.5 speculates on why Samoan and Māori may have undergone change towards

different default allomorphs, despite presumably starting with very similar phonological systems and stem phonotactics.

5.1 Background: Māori phonology

Māori, also known as *te reo* ('the language'), is spoken by the Māori people of mainland New Zealand. The number of Māori speakers has declined rapidly since 1945, though rates of attrition have been slowed by language revitalization efforts (Jones 2012). The 2018 New Zealand census reported that about 186,000 people, or 4.0% of the New Zealand population, could hold a conversation in Māori (Stats NZ 2018). Notably, there are virtually no native speakers of Māori left, excluding some isolated regions. The Māori spoken by the majority of speakers is a second language learned from textbooks.

Māori falls into two major dialect groups, North Island and South Island, the latter of which is extinct (Biggs 1989). The North Island dialects generally have the same segmental phonology, but can further be divided into two dialect groups, Eastern and Western. Beyond this, the dialectal divisions of Māori are still not well understood (Biggs 1989).

Both Māori and Samoan belong to the Nuclear Polynesian language family. Their relationship is visualized in Fig. 5.1, which shows the standard subgrouping of Polynesian languages, with Samoan falling under Proto-Samoic-Outlier and Māori falling under Proto-Eastern-Polynesian (Pawley 1966, 1967; Green 1966).

More recent work has called for the revised subgrouping given in Fig. 5.2. Here, Tongic (Tongan and Niuean) remains the first group to have diverged from the other languages (Nuclear Polynesian). However, the Samoic-Outlier group is abandoned and instead, both Samoan and Māori fall under Proto-Ellician, with Māori being further subgrouped under Proto-Eastern Polynesian (Wilson 1985, 2012; Marck 1996, 1999, 2000, etc.).

Māori has been documented in various work, starting with missionary documents in
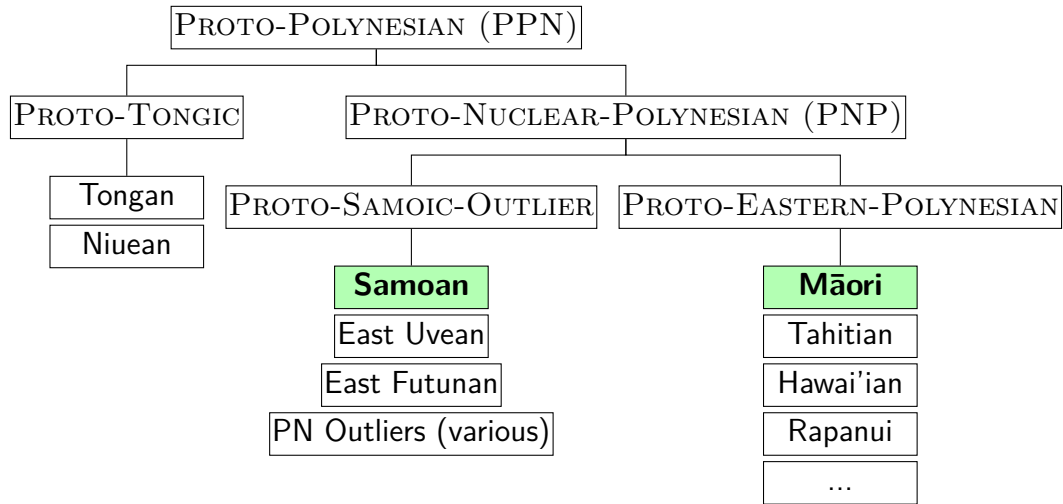
Figure 5.1: The standard subgrouping of Polynesian languages
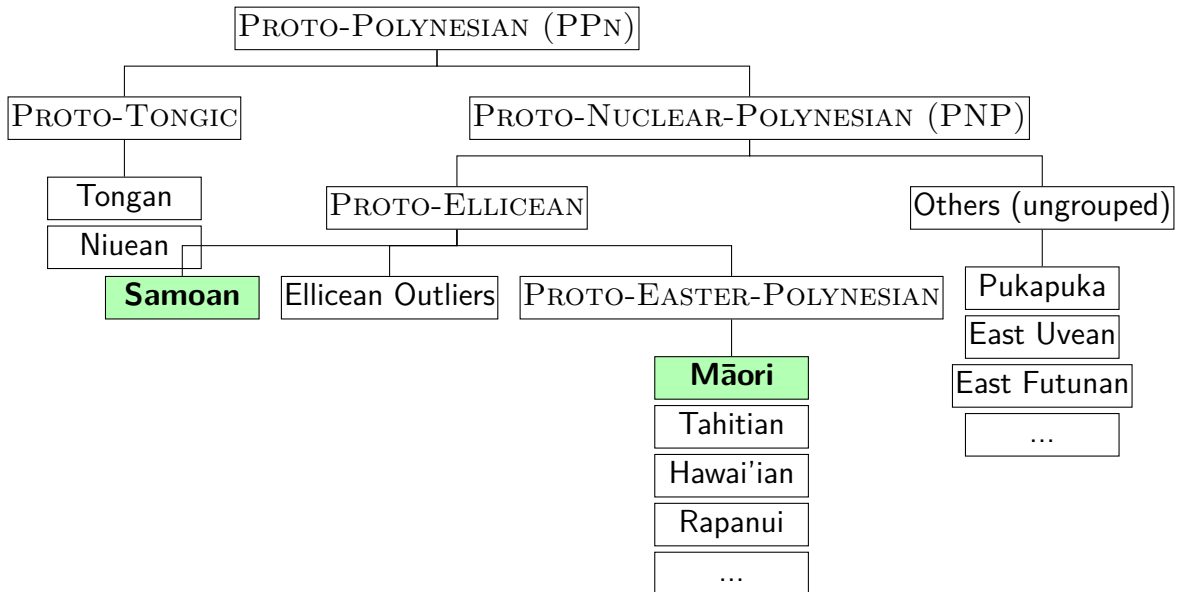
Figure 5.2: The revised subgrouping of Polynesian languages (Marck 2000)

the 19th century. William Williams's (1844) *A Dictionary of the Maori Language* was the first complete dictionary of the Māori language. Since then, the dictionary has been revised multiple times, most recently as the 7th edition by Herbert Williams (1971). Other dictionaries include those by Biggs (2013), Ngata (1971), and Ryan (2012). Grammatical descriptions of Māori include Hohepa (1967), Biggs (1961), Bauer (1993, 1997), and Harlow (2007). As described in Chapter 4, the historical subgrouping of Polynesian lan-

136

guages has also been studied in detail (e.g. Dempwolff 1929; Pawley 1966, 1967; Clark 1973; Greenhill & Clark 2011).

The Māori passive in particular has received much theoretical attention, and is the focus of most debates around the Polynesian thematic consonants. This body of work includes: Hohepa (1967), Hale (1968, 1973), McCarthy (1981), Sanders (1990, 1991), Blevins (1994), Kibre (1998), de Lacy (2003), and Jones (2008). The focus of these studies have generally been on the representation of thematic consonants in synchronic Māori, so the patterns of reanalysis in Māori thematic consonants are still not well understood.

In the rest of this chapter, basic descriptive facts about Māori phonology will be taken from Harlow (2007). Additionally, I draw on Hale (1973), Blevins (1994), and de Lacy (2003) for generalizations about the synchronic distribution of passive allomorphs in Māori. Finally, all examples come from the Williams 7th edition dictionary (Williams 1971), which covers a range of North Island dialects, but is primarily based off of speakers from Ngāpuhi, which falls under the Western dialect region.

5.1.1   Phoneme inventory and phonotactics

Māori has a (C)V(V) syllable structure; codas and final consonants are not allowed, and the rhyme can be either a short vowel, long vowel, or diphthong. Vowels hiatus is also allowed (e.g. [ta:.no.a] 'to belittle'). The status of stress in Māori is controversial and speakers do not have strong stress judgments (Harlow 2007), but most analyses agree that stress is assigned on the leftmost heavy syllable (where a heavy syllable has a diphthong or long vowel nucleus). Otherwise, all else being equal, stress is assigned on the leftmost syllable.

Māori has five phonemic vowel qualities /a e i o u/, all of which contrast in length. Additionally, diphthongs are any sequence of two vowels where the second vowel is higher (/ai, ae, ao, au, ei, eu, oi, ou/). In some dialects, vowel sequences of equal sonority (e.g. [eo], [ui]) also count as diphthongs (Harlow 2007; de Lacy 2003). Other vowel sequences

(e.g., [oa]) form separate syllables.

The Māori consonant inventory is given in (52). Note that /w/ is contrastive with /u/, as evidenced by minimal pairs like [uara] 'desire, value' and [wara] 'murmur'. The labial fricative phoneme is written as /f/ here, but has variable phonetic realization, and is typically realized as [f] or [ɸ]. Finally, the alveolar tap phoneme /ɾ/ is written as <r> throughout the rest of this chapter.

In terms of stem phonotactics, there are several restrictions motivated by local rounding dissimilation. First, the sequences /fo, fu, wo, wu/ (i.e. labial consonants before round vowels) only exist in a few English borrowings. Similarly, /uo/ sequences are never observed except in a few borrowings (Harlow 2007).

(52)   *Māori consonant inventory*

LABIAL	ALVEOLAR	VELAR	GLOTTAL
p	t	k	
f(~ɸ)			h
m	n	ŋ	
	ɾ <r>	w	

5.1.2 The Māori passive and thematic consonant alternations

In Māori, thematic consonant alternations are observed in two suffixes, the nominative (/-ŋa, -Caŋa/) and the passive. Because the nominative suffix is less frequent and not fully productive, I focus on the passive suffix.

The passive suffix has allomorphs /-a/, /-ia/, /-Cia/, /-ina/, and /-na/. The Williams dictionary (1971) contains occurrences of the following /Cia/ passive allomorphs: /-hia, -kia, -mia, -ŋia, -ria, -tia/; examples of each are given in Table 5.1. Of all the passive allomorphs, /-a/ and /-tia/ are the most frequent and often analyzed as the 'default'

138

passives. Note also that in some dialects of Māori, the default allomorph is /-hia/ instead of /-tia/ (Blevins 1994).

de Lacy (2003) analyzes the distribution of /-a/, /-ia/, and /-tia/ as phonologically predictable, and Jones (2008) similarly finds that their distribution correlates with factors such as identity of the stem-final vowel and prosodic shape of the stem. However, as will be discussed in §5.3, /-tia/ is often observed where we might instead expect /-ia/.

Three other allomorphs, /-hina, -fia, -nia/, are extremely rare and were each observed for 1-2 words. In all cases, the stems were also listed with variants that took regularly occurring allomorphs (e.g. [roko-hina]∼[roko-hia] 'to be found').

allomorph	stem	suffixed	POc	gloss
a/ia	fao	fao-a	*paRo	'perforate, chisel'
	paː	paː-ia	*paRa	'stockade'
mia	inu	inu-mia	*inum	'to drink'
tia	ai	ai-tia	*qait	'to copulate'
na/ina	aŋi	aŋi-na	*aŋin	'to blow'
	uta	uta-ina	*qutan	'interior, inland'
ria	mataku	mataku-ria	*matakut	'to be feared'[a]
kia	rere	rere-kia	*rere	'carried by wind'[a]
ŋia	kuː	kuː-ŋia	*guRuŋ	'to coo'
hia	motu	motu-hia	*motus	'to separate'

[a]In these forms, this POc final consonant actually does *not* match the modern one, indicating reanalysis.

Table 5.1: Passive suffix allomorphy in Maori

The passive suffix descends from the Proto-Polynesian CIA suffix. Allomorphy of this CIA suffix (between /-a/, /-ia/, /-ina/, /-na/, /-Cia/) dates back to regular sound changes between POc and PPn (Pawley 1962; Evans 2001); these changes were introduced in Chapter 4, but are given again here and summarized in (53).

In POc, the ergative was originally two suffixes: the short transitive marker *-i followed by the third person pronominal clitic, which was realized as *-a and sometimes *-na (e.g. POc *kila-i-a 'know it'). When *-a/-na ceased to be productive, the suffixes were reanalyzed as a single unit *ia/ina, corresponding to the Māori passive. In Māori, /-ina/ is relatively infrequent and observed mostly as a reflex of POc stems that ended in

*n (discussed below).

Additionally, the POc transitive *-i had phonologically conditioned allomorphy, and was deleted after stems ending in *i and *e (e.g. POc *kani-∅-a 'eat it' vs. *kila-i-a 'know it'). *i is potentially also deleted after *o and *u (Pawley 2001), but the evidence for this is less clear. Either way, *i-deletion derives the /-a/ passive allomorph, which regularly surfaces after most historically vowel-final stems.

Following *n-final POc words, the observed suffix allomorphs are /-ina/ and /-na/, rather than /nia/. /-ina/ arose by metathesis from pre-Polynesian *-nia, mainly when the verb base ended in *a. /-na/ likely also arose by metathesis of *-nia to *ina, followed by deletion of the *i vowel.

(53) *Development of Māori ergative suffix*

Suff.	/ia/	/a/	/Cia/	/ina/	/na/	
POc	*nofo-i-a	*tari-i-a	*bikit-i-a	*koran-i-a	*aŋin-i-a	CHANGE
	-	talia	-	-	-	*i-deletionᴬ
	-	-		koraina	aŋiina	Metathesisᴮ
	-	-			aŋina	*ina -> na
Mao.	noho-ia	(ta)tari-a	piki-tia	(faka)kora-ina	aŋi-na	
	'sit'	'wait'	'pressed close'	'to fuel'	'to blow'	

A. *i-deletion: *i is deleted after *i,*e. **Note:** Evans (2001) argues that deletion of *i happens after stems ending in all vowels other than *a, but Pawley (2001) proposes that the evidence for deletion after *o and *u is less conclusive, citing forms like [nofo-ia] (<*nofo) 'to sit', where *i is maintained after the back vowels.

B. Metathesis of *nia to *ina after *a

In summary, PPn *-a, *-ia and *-ina correspond to stems that were vowel-final in POc, with the additional wrinkle that *-ina is homophonous with the allomorph that occurs after stems ending in *an. In Samoan, /-a/ and /-ina/ were passed down and are the productive allomorphs of the ergative suffix, while /-ia/ is relatively rare. In contrast, for Māori, of the allomorphs corresponding to vowel-final stems, /-a/ and /-ia/ are observed while /-ina/ is relatively uncommon.

Table 5.2 summarizes the regular sound correspondences between POc and Māori. Based on these, we can infer the passive allomorph that should surface based on the POc final segment. Notably, POc stems ending in *p, *pw, and *s are all expected to take /-hia/ in Māori as a result of two sound changes. First, PPn *s became Māori [h] in all environments. The changes that affected PPn *f (i.e. POc *p/pw) are more complicated; in general, *f stayed as [f] word-initially before rounded vowels ([u, o]), and changed to [h] elsewhere. Based on this distribution, stem-final *f, which is intervocalic in suffixed forms, should reflect as /-hia/. However, Harlow (2007) notes that there is some variation in how PPn *f is reflected, where [h] is sometimes observed in the environments where [f] should surface and vice-versa. In modern Māori, /-fia/ is almost never observed (N = 2), so I assume that stem-final PPn *f always reflects as /-hia/.

POc	PPn	Mao	Passive
*p,*pw	*f	f	hia[2]
*t, *j, *d	*t	t	tia
*l, *r, *dr	*l, *r	r	ria
*k, *g	*k	k	kia
*m	*m	m	mia
*n, *ñ	*n	n	na, ina
*ŋ, *mw	*ŋ	ŋ	ŋia
*s	*s	h	hia
*c	*h	∅	(i)a
*q	*ʔ	∅	(i)a
*y, *R	∅	∅	(i)a

[1]POc *b/*bw and *w, whose Māori reflexes are [p] and [w], are excluded here because they are not found word-finally in POc, and therefore never reflect as thematic consonants.
[2]In Māori, PPn *p generally became [f] word-initially before unrounded vowels, and [h] elsewhere.

Table 5.2: Māori reflexes of POc final consonants[1]

5.1.3 Predictable allomorphy

Although the Māori passive has many allomorphs, the relative distribution of some of them is phonologically predictable. In this section, I discuss these generalizations, drawing on de Lacy's (2003) analysis of the Māori passive. All generalizations are based on

a set of 1167 stem-passive pairs (de Lacy, p.c.), which are primarily sourced from the Williams 7th edition dictionary and supplemented with data from de Lacy's fieldwork with the Ngāti Pōrou and Ngāti Awa tribes (who speak North Island dialects).

First, the relative distribution of /-na/ and /-ina/ is predictable, with /-ina/ surfacing after [a]-final stems and /-na/ surfacing elsewhere (e.g. [tipako-na] 'to select' vs. [kata-ina] 'to laugh at'). Of the 80 stems that take either /-na/ or /-ina/, there is only one exception to this generalization, [eke]∼[eke-ina] 'to climb' (cf. *[eke-na]).

The relative distribution of /-a/ and /-ia/ is also mostly predictable, with /-ia/ surfacing after [a]-final stems and /-a/ surfacing elsewhere (e.g. [apu-a] 'cover' vs. [tapa-ia] 'recite'). As noted above, this distribution is thought to date back to allomorphy already present in POc (Evans 2001; Pawley 2001). de Lacy (2002, 2003) analyzes the distribution of /-a/ vs. /-ia/ as motivated by a constraint *OCP-V, which penalizes adjacent identical vowels across morpheme boundaries.

Around 5.6% of stems that take /-ia/ or /-a/ (n = 30/534) do not follow this generalization; as seen in Fig. 5.3 below, most exceptions involve stems of the type [uŋa]∼[uŋa-a] 'to send', where an [a]-final stem is followed by the passive allomorph /-a/.
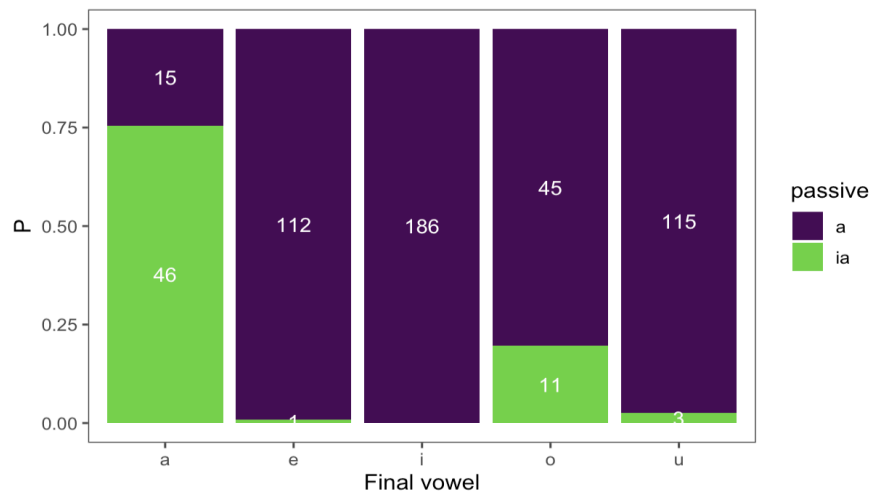
Figure 5.3: Distribution of /-a/ and /-ia/ by stem-final vowel

142

5.1.4 Distribution of /-(i)a/ and /-tia/

The relative distribution of /-(i)a/ and /-tia/ is partially predictable from several phono-logical factors (that are incidentally similar to the ones present in Samoan suffix allomor-phy): identity of the stem-final vowel, quantity, and prosodic shape. A formal analysis of these patterns is given in de Lacy (2003). In this section, I take a different approach and focus on surface distributions, where they might be relevant to reanalysis.

First, as already discussed in the previous section, the distribution of /-a/ and /-ia/ is predictable from the identity of the stem-final vowel, in a way that is transparently motivated by an OCP-V constraint. Additionally, as noted by Blevins (1994), the passive which surfaces is sensitive to the quantity of the base of suffixation. Fig. 5.4 shows the distribution of suffixes by the number of moras in the base. As a word increases in length, the tendency to take /-tia/ as a suffix also increases. The corpus I use does not include recent borrowings, but Blevins (1994) further notes that in longer loanwords, /-tia/ is always the suffix that surfaces.
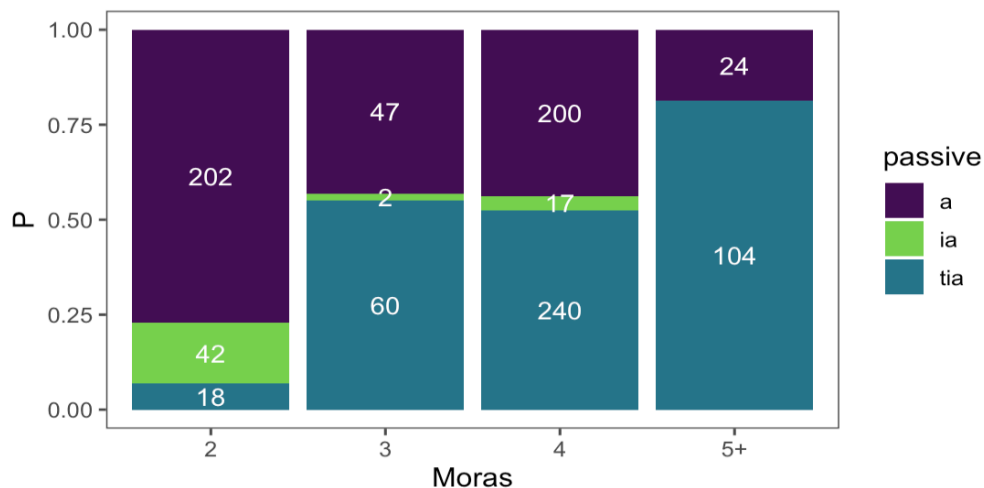
Figure 5.4: Distribution of /-a/, /-ia/, and /-tia/ by number of moras in stem

Finally, prosodic shape influences the passive which surfaces. Fig. 5.5 shows the distribution of /-a/, /-ia/, and /-tia/ by prosodic shape. This figure includes the subset of forms that are under five moras. H(eavy) syllables have a long vowel or diphthong

143

nucleus, while L(ight) syllables have short vowel nuclei. Additionally, [a]-final stems are presented separately because they behave differently from other stems; that is, they take /-ia/ instead of /-a/ due to avoidance of OCP-V.

For the [a]-final stems, there is a generally greater tendency to take /-tia/ instead of /-ia/. Notably, analyses of passive allormorphy in Māori generally predict that all bimoraic stems (i.e. H and LL stems) should take /-a/ or /-ia/ as the passive suffix. We also see some difference between the [a]-final stems and other stems; when a bimoraic stem is [a]-final, it is slightly more likely to take /-tia/.

For the longer stems (i.e. 3+ moras), there is more variation in the suffix that is selected. This could be because most longer words are either prefixed or reduplicated, but there is lexical variation in whether this extra morphological material is fossilized (and analyzed as belonging in the same prosodic word as the root), or still parsed into separate morphemes (belonging to separate prosodic words).
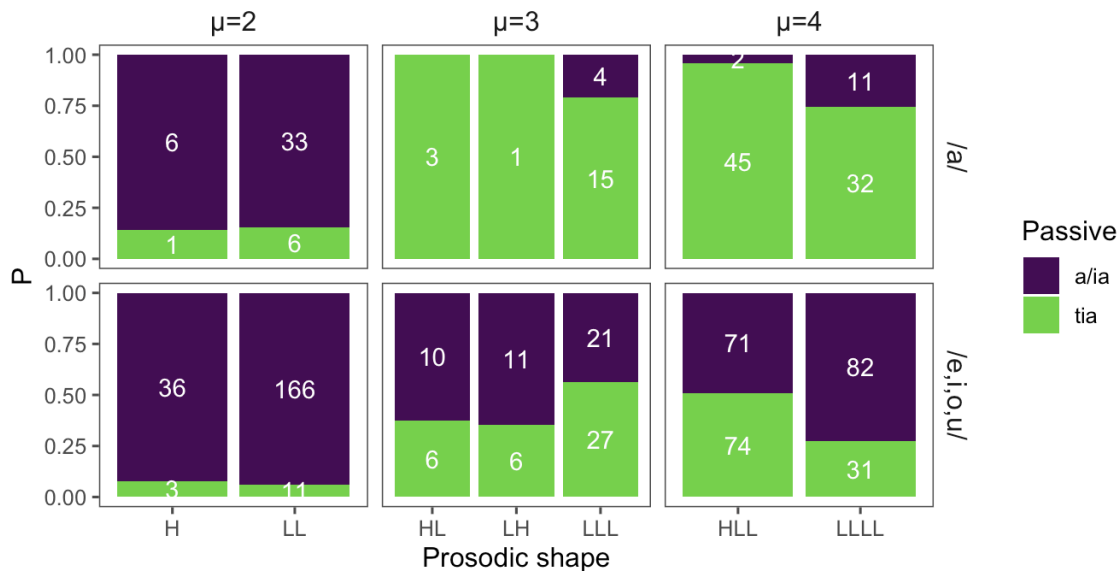
Figure 5.5: Distribution of /-(i)a/ and /-tia/ by prosodic shape

Note that de Lacy (2003) treats stems which take a /-Cia/ allomorph as underlyingly consonant-final (e.g. [inum]~[inumia] /inum-ia/ 'to drink'). Under this analysis (which recapitulates the historical development of thematic consonants), consonant-final stems

always take a predictable passive allomorph /-ia/. However, if we consider forms that take a /-Cia/ allomorph, it turns out that their distribution is asymmetrical.

Fig. 5.6 shows the distribution of /-(i)a/, /-tia/, and /-Cia/ by prosodic shape; it is essentially the same figure as Fig. 5.5, with the addition of /-Cia/ (where /-Cia/ is any allomorph beginning with a consonant other than /t/). In the bimoraic forms, there is a discrepancy between the [a]-final stems and other stems, where [a]-final stems are more likely to take /-Cia/ compared to other stems. Overall, [a]-final stems generally take a higher proportion of consonant-initial passive allomorphs, including both /-tia/ and other /-Cia/ allomorphs. This point becomes important in §5.3 below.
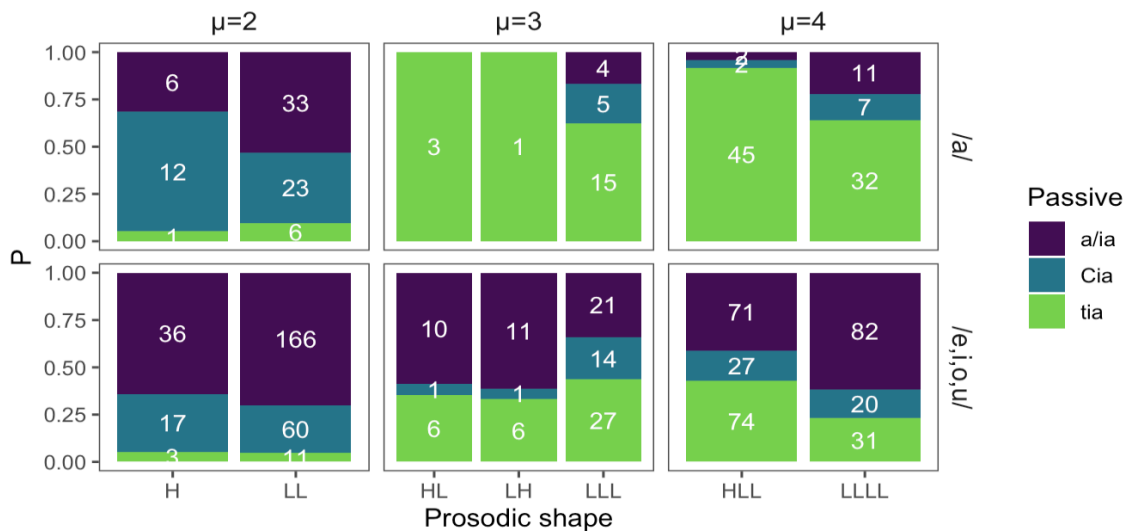
Figure 5.6: Distribution of /-(i)a, /-tia/, and /-Cia/ by prosodic shape

5.1.5 OCP-place avoidance and /-tia/

Recall that in Samoan, the distribution of /-Cia/ allomorphs was conditioned by OCP-place constraints. In particular, there was a strong effect of OCP-COR-SON, which penalizes sequences of coronals that share the same sonorancy specification (e.g. *[ra**n**u] and *[**t**a**t**u], but not [**t**a**n**u]).

/-tia/ has a much more general distribution in Māori than in Samoan. Despite this, the

distribution of /-tia/ in Māori appears to be somewhat sensitive to OCP-place restrictions. In particular, /-tia/ occurs at a lower proportion when the nearest preceding consonant is a /t/; Māori does not have /s/ owing to the historical change of PPn*s>[h] , so /t/ is the only consonant that could trigger OCP-COR-SON effects when the passive allomorph is /-tia/.

This OCP-COR-SON effect is demonstrated in Fig. 5.7. The leftmost subfigure shows the proportion of passive allomorphs in the lexicon, separated by whether or not the nearest preceding consonant is /t/. When the stem has a preceding /t/, /-tia/ occurs at a lower proportion.

Importantly, /-tia/ is less frequent specifically when the preceding consonant is /t/, and not just when the stem contains any preceding consonant. To demonstrate this, two other subfigures are shown in Fig. 5.7; these respectively group stems by whether they have a preceding coronal sonorant (/n, r/) , and whether they have a preceding labial (/p, f, m, w/). In both figures, the proportion of stems that take /-tia/ is relatively uniform regardless of the preceding consonant.
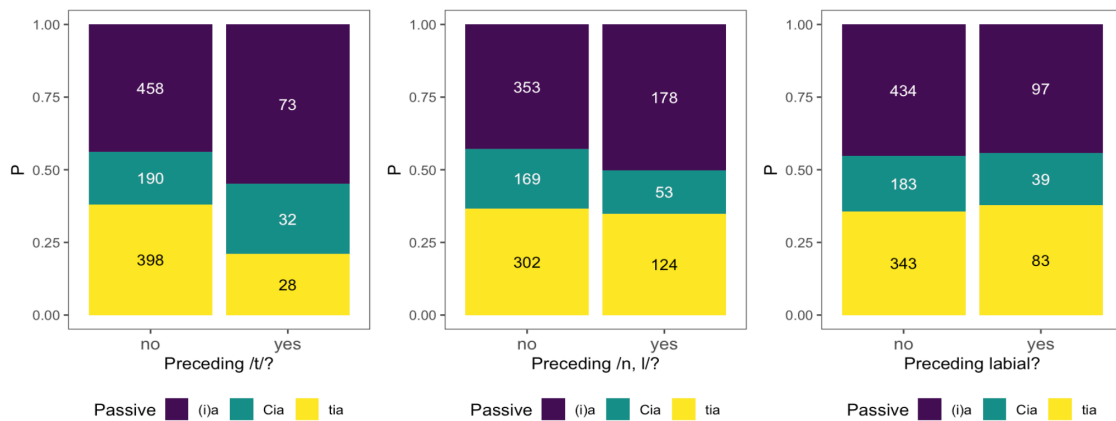
Figure 5.7: Distribution of /-(i)a/, /-tia/, and /-Cia/ by preceding consonant of stem

5.2 Stem phonotactics

In this section, I discuss the stem phonotactics argued to be the basis of reanalysis in Māori. As a preview, Māori is shown to have a dispreference for vowel hiatus, particularly when hiatus involves long vowels and diphthongs. I argue that this pattern can be explained as the cumulative effect of two constraints, which respectively penalize vowel hiatus and heavy syllables (i.e. ones with long vowels or diphthongs).

5.2.1 Vowel hiatus and correption

The examples in (54) demonstrate how constraints against hiatus and heavy syllables can motivate reanalysis towards /-tia/ (or more generally any consonant-initial allomorph). For now, I use *HIATUS to penalize vowel hiatus, and *LONGNUCLEUS (henceforth *LONGNUC) to penalize long vowels and diphthongs; options for how to enforce hiatus are discussed below.

When an [a]-final stem takes passive /-a/, the resulting stem, exemplified by (54c), violates OCP-V. Violations of OCP-V are very rare in Māori, and this option is presumably the most marked. This leaves us with the other three possibilities.

When an [a]-final stem takes /-ia/ as the passive allomorph, the resulting suffixed form violates both *HIATUS and *LONGNUC; an example of this is given in (54a). Note that in words like /faka-ia/, the resulting /a-i/ sequence is syllabified together as a diphthong (i.e. [fa.kai.a], not *[faka.i.a]) (de Lacy 2003; Harlow 2007). In contrast, [a]-final stems which take /-tia/ violate *HIATUS but not *LONGNUC; all other /-Cia/ allomorphs would have the same violation profile as /-tia/. In (54d), the stem which takes /-ina/ avoids violations of *HIATUS. However, the resulting suffixed form has a diphthong [ai], and therefore incurs a violation of *LONGNUC. Importantly, because stems which take /-ia/ violate both *HIATUS and *LONGNUC, they are more marked than stems that take either /-ina/ or /-tia/.

(54)   *Hiatus in passive forms of [a]-final stems*

		STEM	SUFFIXED	GLOSS	MARKEDNESS
a.	/STEM-tia/	[wa.ha]	[wa.ha.ti.a]	'raise up'	*HIATUS
b.	/STEM-ia/	[ho.ka]	[ho.kai.a]	'run out'	*HIATUS, *LONGNUC
c.	/STEM-a/	[pa.na]	[pa.na.a]	'expel'	*HIATUS, OCP-V
d.	/STEM-ina/	[ka.ta]	[ka.tai.na]	'laugh at'	*LONGNUC

In stems that are *not* [a]-final, the allomorph that surfaces is typically /-a/ rather than /-ia/. As demonstrated in (55), this means that all suffixed forms incur a violation of *HIATUS. All else being equal, there isn't a markedness motivation to prefer one allomorph over another. These generalizations, taken together, mean that if reanalysis is motivated by *HIATUS and *LONGNUC, it should target the [a]-final stems that take /-ia/. As discussed in the following section, this turns out to be the case.

(55)   *Hiatus in passive forms of stems that do not end in /a/*

		STEM	SUFFIXED	GLOSS	MARKEDNESS
a.	/STEM-a/	[pe.pe]	[pe.pe.a]	'to flutter'	*HIATUS
b.	/STEM-tia/	[pi.ki]	[pi.ki.ti.a]	'downwards'	*HIATUS

The marked status of vowel hiatus is well-substantiated. Hiatus avoidance is widespread (e.g. Casali 1997; Siptár 2003), and often argued to be aimed at removing onsetless syllables (Blevins 1995). In Optimality Theory, hiatus avoidance is typically enforced with a constraint NOONSET (McCarthy & Prince 1993, 34-37).

Hiatus specifically involving long vowels (in this case VV.V sequences) is less well-understood than general hiatus effects, but there is some evidence that VV.V hiatus behaves differently from hiatus of two short vowels. In particular, correption, or the prevocalic shortening of long vowels, has been observed in various languages like Latin (Mester 1994), Greek (Sihler 1995, 74), Hungarian (Siptár & Törkenczy 2000, 125-128), Korean (Kim 2000, 61-64), and Kikamba (Roberts-Kohno 2000). Correption is also well-attested in poetic meter (e.g. Kiparsky 1989; Clapp 1906; Hanson 2001; Gunkel & Ryan 2011).

There is also cross-linguistic evidence for the markedness of long vowels and diphthongs. Diphthongs, in particular, are representationally complex and face conflicting demands between maximizing perceptual distinctiveness and minimizing articulatory effort (Minkova & Stockwell 2003; Flemming 2004; Petersen 2016).

As mentioned above, I argue that for Māori, the dispreference for VV.V sequences can be modeled as the cumulative effect of two intersecting constraints, a constraint against long vowels/diphthongs, and a constraint against hiatus. The following section uses a quantitative account of Māori stem phonotactics to substantiate this claim.

5.2.2 Hiatus avoidance and *LongNuc in Māori phonotactics

Generalizations made in this section are all based on 7430 headwords taken from the Williams 6th edition dictionary (Williams 1957). In pre-processing, forms were removed if they took one of the common prefixes (/kai-/ 'agentive noun', /ma-/ 'adjective', /faka-/ 'causative', /tua-/ 'ordinal'), if they were reduplicated, or if they were clearly compounded. Patterns were also confirmed separately with a list of 1000 frequent words taken from the Ministry of Education (Te Kete Ipurangi). This list was based on two Māori corpora: the Corpus of Māori Texts for Children, compiled by Huia Publishers, and the Māori Broadcast Corpus (MBC; Boyce 2006).

Table 5.3 shows the distribution of vowel nuclei in the Williams corpus. Monophthongs account for almost 90% of the lexicon and are evidently much more frequent than both long vowels and diphthongs. This general dispreference for long vowels and diphthongs can be enforced using the constraint *LongNuc.

Type	Count	P
Monophthong	6047	0.89
Long vowel	248	0.04
Diphthong	454	0.07

Table 5.3: Distribution of vowel nuclei in the Williams corpus

Fig. 5.8 shows the distribution of syllable-syllable pairs in Māori (using the Williams

corpus), and in particular demonstrates the effects of both *LONGNUC and the dispreference for hiatus. Forms are grouped by whether syllables were light (V) or heavy (VV), and by whether or not there was vowel hiatus at the syllable boundary. For example, [ke.a] 'phlegm' contains a V.V sequence, while [aː.he.a] has both a VV.CV sequence and a V.V sequence. Both diphthongs and long vowels are represented as VV.
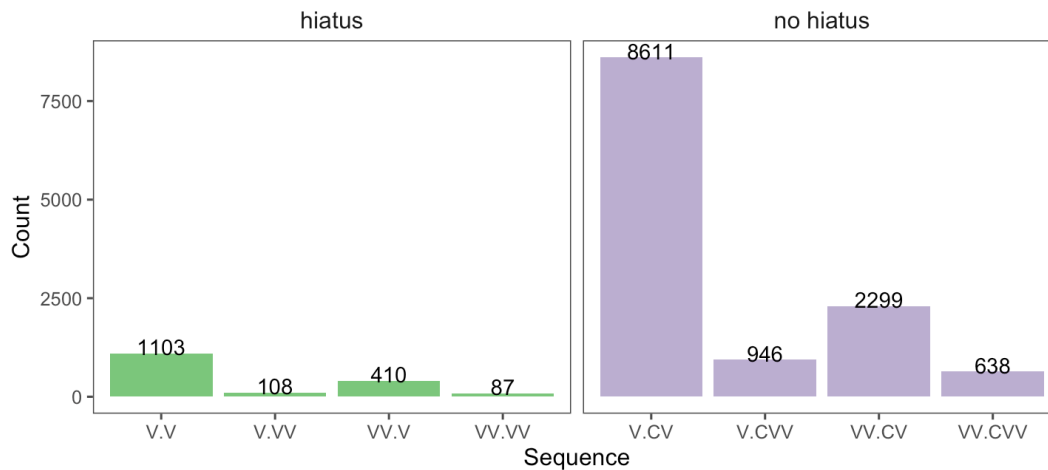
Figure 5.8: Counts of syllable-syllable combinations in the Williams dictionary

One clear pattern that can be observed here is the dispreference for hiatus; syllable-syllable sequences with hiatus (shown in the left-hand figure) are overall much lower in frequency. Within each figure, we also see that syllable-syllable sequences which include long nuclei are generally dispreferred. This pattern is observed for both the 'hiatus' and 'no hiatus' conditions, suggesting that it may be an effect of the generally low frequency of VV syllables.

Interestingly, V.(C)VV sequences (i.e. short vowels followed by long vowels) are also lower in frequency than VV.(C)V sequences. This pattern is observed in both the 'hiatus' and 'no hiatus' conditions, suggesting that it is not specific to hiatus. Nevertheless, it should be noted that dispreference for V.VV relative to VV.V sequences is typologically unusual, as in the literature, a dispreference for VV.V (motivating correption) is more well-attested.

These patterns are once again confirmed using a MaxEnt phonotactic model, following

the procedure laid out by Wilson & Obdeyn (2009). The phonotactic grammar shown here is relatively simple and aimed at illustrating the effect of hiatus and syllable weight in Māori phonotactics. Subsequent modeling results (§5.4.2) are based on a more nuanced phonotactic grammar that accounts for other constraints (Hayes & Wilson 2008).

Model input counts were derived from the same corpus of 7430 headwords used above, taken from the Williams 6th edition dictionary. Words with five or more syllables were excluded, leaving 6939 items. Additionally, inputs were simplified to reflect only their syllable structure. For example, a stem like [apai] is coded as 'V.CVV'. The input was all possible stem shapes with four or less syllables, where each syllable can vary in terms of whether it has an onset (C vs. $\varnothing$), and whether it has a heavy or light nucleus (V vs. VV).

The constraint set includes singleton constraints penalizing each nucleus type (i.e. *SHORTNUC and *LONGNUC). Additionally, I tested four constraints for hiatus: *HIATUS, NOONSET, *VV-V, and *V-VV. As discussed above, hiatus is often argued to be motivated by avoidance of onsetless syllables. To see whether this is the case for Māori, I test both HIATUS and NOONSET.

*VV-V targets the environment for correption, while *V-VV targets the opposite hiatus environment. I also include two constraints *VV-(C)V and *V-(c)VV which penalize the respective syllable-syllable combinations in both hiatus and non-hiatus environments. These constraints are included because, as seen in Fig. 5.8, V.(C)VV sequences appear to be dispreferred regardless of whether or not there is hiatus.

Finally, longer words are generally less frequent than shorter words. The model could learn spuriously high weights for *SHORTNUC and *LONGNUC to account for the lower frequency of long words. To prevent this, a constraint *STRUC-μ is included; this constraint assigns a violation for every mora in a word, and therefore penalizes long words.

Just like in previous chapters, constraints were tested for significance using Likelihood Ratio Tests (Hayes et al. 2012). The results of these tests are given in Table 5.4.

NOONSET was found to be highly significant (as evidenced by the large change in log

Constraint	w	ΔL	p
*Struc	0.15	199.1	$<1\times10^{-15}$
*ShortNuc	0	0.04	*n.s.* (0.84)
*LongNuc	0.61	408.0	$<1\times10^{-15}$
NoOnset	2.00	1285.2	$<1\times10^{-15}$
Hiatus	0	0	*n.s.* (1)
VV.V	0	0	*n.s.* (1)
V.VV	0.12	2.47	*n.s.* (0.12)
VV.(C)V	0	0	*n.s.* (1)
V.(C)VV	1.60	1043.1	$<1\times10^{-15}$

Table 5.4: Likelihood Ratio Test results for a model of Māori syllable phonotactics

likelihood). On the other hand, Hiatus was non-significant and assigned zero weight by the model. These results suggest both that hiatus avoidance is active in Māori stem phonotactics, and that this behavior is best characterized as avoidance of onsetless syllables.

Looking at the other constraints, *LongNuc was also found to be strongly significant, while *VV.V and *VV.(C)V were both non-significant. Taken together, these results suggest that Māori's dispreference for VV.V sequences is best characterized as the effect of two interacting constraints (NoOnset and *LongNuc), rather than the effect of a constraint that specifically targets the environment for correption.

Interestingly, *V.(C)VV was also found to be significant, while *V.VV was not. In other words, the dispreference for light-heavy syllable sequences is best captured by a constraint that penalizes both hiatus and non-hiatus environments. One possible explanation for this is that V.VV and V.CVV sequences often result in a conflict of stress assignment constraints. Recall that in Māori, stress is weight-sensitive (falling on the heaviest syllable), but otherwise aligned to the left edge of a word. Words of the shape V.VV, such as [itáu] 'girdle', will therefore incur violations of a constraint that requires left-alignment.

5.3 Patterns of reanalysis in Māori

In this section, I look at the distribution of final consonants in POc and compare this to modern Māori, in order to probe at the patterns of reanalysis in Māori weak stems. The POc data is the same set of 1,023 protoforms used for Samoan; forms were sourced from the ACD (Blust & Trussel 2010) and must have at least 6 cognates within the Oceanic language family. The Māori data is the same corpus of 1167 stem-passive pairs used in §5.1 above (de Lacy, p.c.).

5.3.1 Historical distribution of thematic consonants (POc)

Table 5.5 shows the distribution of final segments in POc, organized by their corresponding passive suffix reflex in Māori. Vowel-final stems, corresponding to /-a/ and /-ia/, were by far the most frequent type of stems. This means that in a frequency-matching model, reanalysis should be towards /-a/ and /-ia/.

/-tia/ and /-hia/ are the most frequent alternants after /(i)a/, each occurring in around 7% of the POc corpus. This is interesting especially since in modern Māori, the default allomorph is generally /-tia/, but also /-hia/ in some dialects (Blevins 2008). Note also that in Māori, the merger of POc *s and *p to [h] means that /-hia/ may be more frequent than in languages like Samoan, where the merger did not happen.

POc	Allomorph	n	P
*vow (or *q, R)	(i)a	672	0.66
*m	mia	20	0.02
*t,j, d	tia	67	0.07
*n, ñ	na/ina	60	0.06
*r,l,dr	ria	36	0.04
*k	kia	52	0.05
*ŋ, mw	ŋia	40	0.04
*s, p	hia	66	0.07

Table 5.5: Distribution of final segments in POc

Even if /-a/ and /-ia/ are treated as separate allomorphs, both are individually more

frequent than the other allomorphs. Recall that /-ia/ is occurs after [a]-final stems, while /-a/ occurs elsewhere. Because /a/ is the most common vowel in POc, /-ia/ is expected to occur very frequently. Specifically, based on the distribution of final vowels in POc, /-a/ would be the most frequent allomorph (n = 347, p = 0.34), closely by /-ia/ (n = 325, p = 0.32).

Fig. 5.9 shows the distribution of passive allomorphs in POc by the identity of the stem-final vowel. /-a/ and /-ia/ are combined into one category here and assumed to be in complementary distribution (with /-ia/ appearing after [a]-final stems). From this figure, we see that /-tia/ is slightly more frequent after /i/-final stems, but otherwise relatively evenly distributed across each vowel. Likewise, the /-Cia/ allomorphs occur roughly evenly across each vowel category. In other words, there is no strong distributional evidence for reanalysis based on identity of the stem-final vowel.
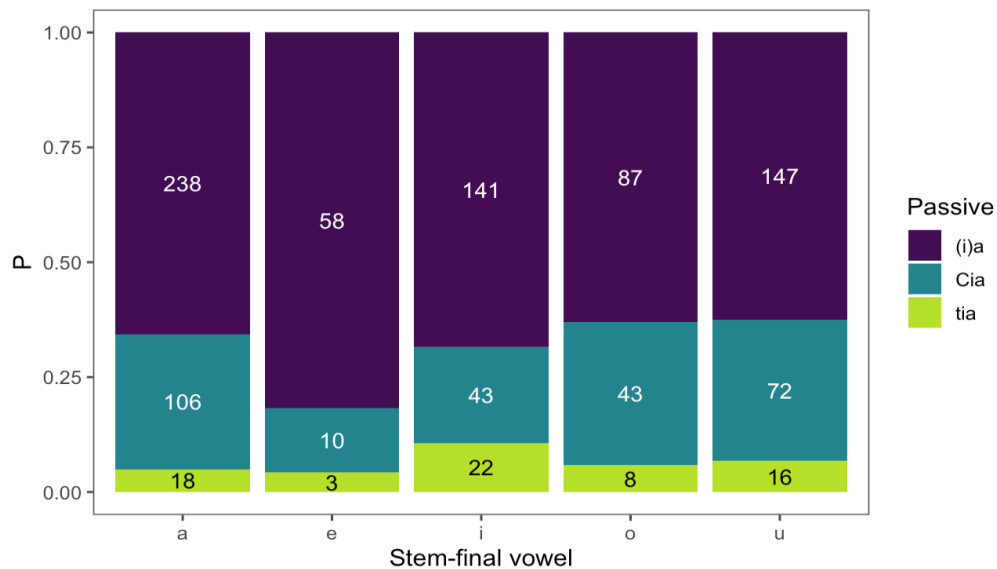

Figure 5.9: Distribution of passive allomorphs by stem-final vowel in POc

5.3.2 Comparison of POc and Māori

In this section, the distribution of passive allomorphs in POc and Māori are compared. This comparison gives us insight into whether reanalysis has occurred towards more fre-

quent alternants, as is predicted by frequency-matching approaches. §5.3.3 gives a form-by-form comparison of observed reanalyses.

Fig. 5.10 compares the overall proportion of allomorphs in POc and Māori; /-na/ and /-ina/ are both grouped together with the other /-Cia/ allomorphs. The proportion of stems taking /-Cia/ allomorphs has decreased between POc and Māori, which is not surprising given the historical POc distribution, where /-Cia/ allomorphs were already the least frequent. However, there has been a decrease in /-(i)a/ and increase in /-tia/ that is not predicted by historical distributions. This discrepancy suggests that there has been reanalysis away from /-(i)a/ and towards /-tia/.
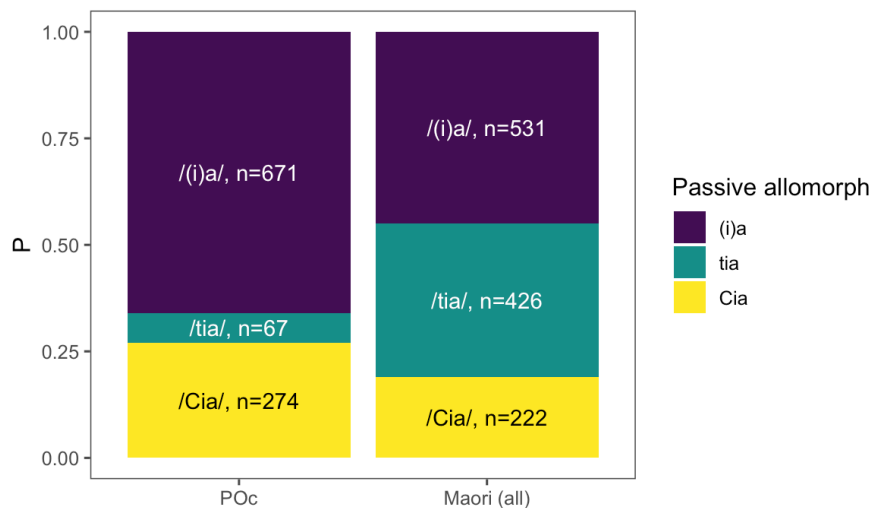
Figure 5.10: Distribution of passive allomorphs in POc vs. Māori

As discussed above, in modern Māori, the distribution of /-(i)a/ and /-tia/ is in part conditioned by the prosodic shape of stems. Generally speaking, bimoraic stems are expected to take /-(i)a/, while longer stems are expected to take /-tia/. Stems of the shape LLLL (with four light syllables) fall somewhere in between.

To see if reanalysis towards /-tia/ has affected the lexicon in general, or just longer words, I also look at the distribution of passive allomorphs in the subset of stems expected to take /-a/ or /-ia/, based on prosodic shape. Fig. 5.11 is identical to Fig. 5.10 above, with the difference that the column titled 'Maori ($\mu < = 4$) shows the subset of Māori

stems expected to prefer /-(i)a/ instead of /-tia/ (i.e. ones with a shape HL, LH, or LLLL). For this subset of stems, the preference for /-tia/ is less skewed, but still present. There is still a decrease in /-(i)a/ compared to POc, suggesting that reanalysis away from /-(i)a/ has happened across the board.
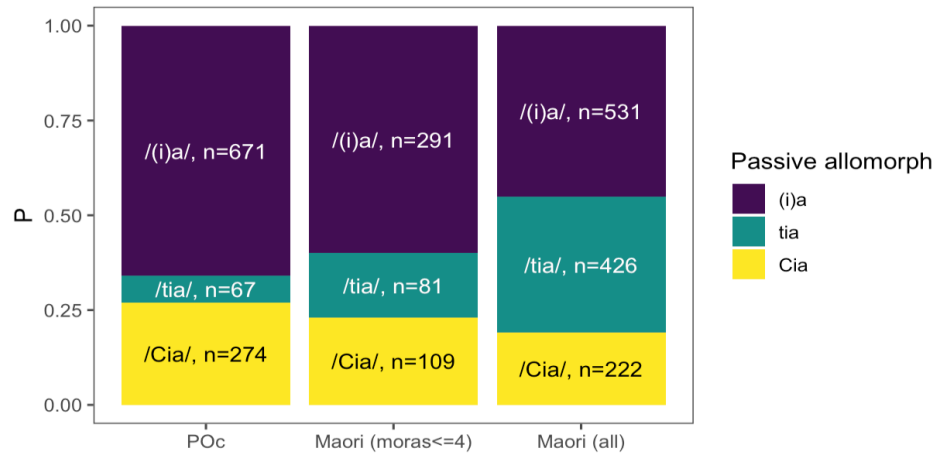
Figure 5.11: Distribution of passive allomorphs in POc vs. Māori, by prosodic shape of stem

Finally, Fig. 5.12 shows the distribution of passive allomorphs by identity of the stem-final vowel. We can observe an effect of the stem-final vowel, such that the shift away from /-(i)a/ has primarily happened in [a]-final stems. Notably, of the consonant-initial allomorphs, /-tia/ is the most frequent (and this is true regardless of the identity of the stem-final vowel). For the /a/ and /o/-final stems, the proportion of consonant-initial allomorphs (i.e. /-Cia/) has also stayed relatively high.

5.3.3 Direct evidence of reanalyses

The results of the above section suggests that there has been reanalysis away from the vowel-initial allomorphs, particularly when the stem-final vowel is /a/. In other words, reanalysis has mostly targeted /-ia/. Recall from §5.2 that this is the direction predicted by a markedness account, if *LONGNUC and NOONSET are active in Māori reanalysis.

In this section, I present case-by-case comparisons of protoforms with their Māori
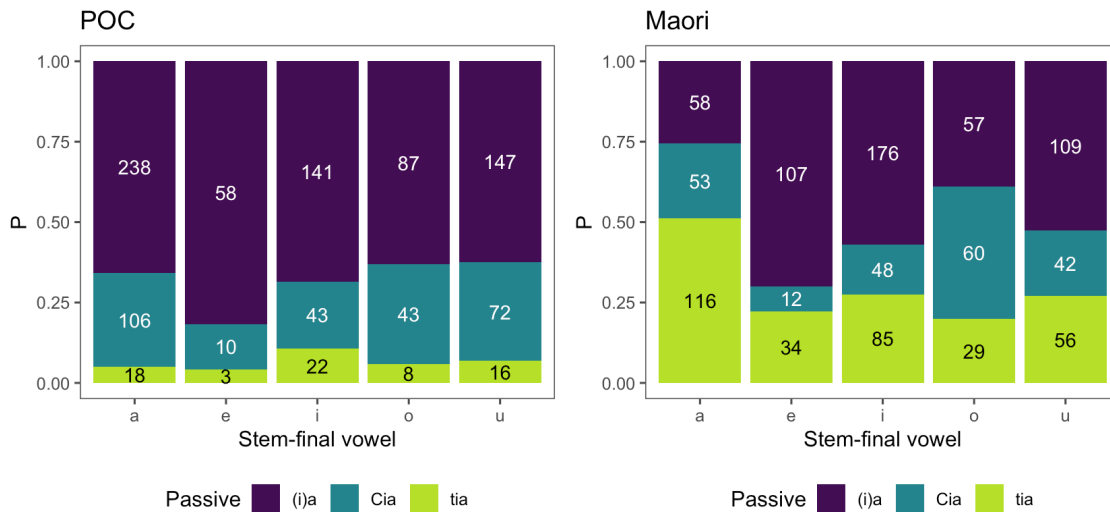
Figure 5.12: Distribution of passive allomorphs in POc vs. Māori by stem-final vowel

reflexes; these comparisons are consistent with the results so far suggesting that reanalysis has targeted /-ia/ more than other allomorphs.

The table in Table 5.6 shows form-by-form comparisons (of POc vs. Māori) for the subset of stems that were expected to take /-a/ or /-ia/. The column named 'match' shows whether the Māori allomorph matches POc (i.e. is /-a/ or /-ia/), or does not match (indicating that reanalysis has occurred). When the expected allomorph is /-ia/, there is a much larger degree of reanalysis; around 63% of forms (n = 19/30) have been reanalyzed, compared to 29% for stems expected to take /-a/ (n = 22/75).

POc	match	n	p
/a/	yes	53	0.71
	no	22	0.29
/ia/	yes	11	0.37
	no	19	0.63

Table 5.6: Mismatches between POc and Māori

157

5.3.4 Interim summary

Overall, results of this section suggest that /-ia/ has been reanalyzed more than would be expected given its high frequency in the historical POc distribution. This discrepancy between the POc and Māori distributions is summarized in Table 5.7.

In general, across vowel contexts, there has an increase in the proportion of words that take /-tia/. However, this difference is particularly striking for the [a]-final stems. Based on the POc distribution, around 66% of stems are expected to take the vowel-initial allomorph /-ia/. In Māori, however, only 22% of [a]-final stems take /-ia/. Instead, most of these stems (57%) have /-tia/ as the passive allomorph.

final V	Passive	POC	Maori
/a/	(i)a	0.66	0.22
/a/	Cia	0.29	0.21
/a/	tia	0.05	0.57
other	(i)a	0.69	0.51
other	Cia	0.24	0.2
other	tia	0.07	0.29

Table 5.7: Summary: distribution of allomorphs in POc vs. Samoan

I propose that reanalysis of /-ia/→/-Cia/ occurred in order to avoid outputs that were phonotactically marked in terms of violating both *LongNuc and NoOnset. Because /-tia/ was the most frequent of the /-Cia/ allomorphs, it was most frequently the result of reanalysis. Over time, as reanalysis made /-tia/ more and more frequent, it overtook the other allomorphs and was extended to more general environments. This resulted in the state of Māori passive allomorphy that we see today, where /-a/ and /-tia/ are the productive allomorphs.

5.4 Modeling reanalysis in Māori

In this section, patterns of reanalysis in Māori are modeled as the result of frequency-matching combined with a markedness bias, specifically against heavy vowels and hiatus.

The model implementation is very similar to what was adopted in Chapter 4 for Samoan. In §5.4.1 and §5.4.2, I discuss points where the model implementation is specific to Māori. §5.4.3 presents model results.

5.4.1 Choice of URs and inputs

As pointed out in Chapter 4, Polynesian thematic consonants can analyzed in two ways. Under the **phonological** analysis, adopted by Sanders (1990, 1991) and de Lacy (2003) among others, the thematic consonant belongs to the stem UR, and the passive suffix has just a few predictable allomorphs. For example, [inu]∼[inumia] 'to drink' can be derived from the URs /inum/ 'to drink' and /ia/ 'PASSIVE'.

Under the **morphological** analysis, which was first proposed by Hale (1968, 1973), the passive has multiple allomorphs, while stems are always underlyingly vowel-final. For example, [inu]∼[inumia] 'to drink' has the URS /inu/ and /-mia/. In the morphological approach, the task of UR learning is simpler because stem SRs and URs are closely matched. On the other hand, morphophonology is more complex, as the grammar now has multiple passive suffix allomorphs with partially unpredictable distribution.

I follow Hale in adopting the morphological approach, for reasons discussed in Chapter 4. Note, however, that the issue of how Māori thematic consonants should be analyzed does not directly affect the question at hand. A model of markedness-driven reanalysis can be implemented regardless of how thematic consonants are represented in the underlying forms.

Examples of an input and its corresponding candidates are given in (56). Following Hale's analysis of Māori, the input stem is vowel-final, and candidates take different suffix allomorphs. In other words, stem URs always match SRs, and the thematic consonants belong to the suffix.

(56)   *Example of candidates for* [inu]∼/inu/-PASS

　　　inu**a**
　　　inu**mia**
　　　inu**tia**
　　　inu**na**
　　　inu**ria**
　　　inu**ŋia**
　　　inu**kia**
　　　inu**hia**

Model inputs are 500 stems whose distribution model that of the POc protoforms. Inputs are pooled by the identity of the stem-final vowel (/a e i o u/) and identity of the immediately preceding consonant (/p,f,m,w,t,n,r,k, ŋ,h/ or 'none'). I do not consider conditioning effects of prosodic shape. Therefore, an input like /ito/ represents all stems where the preceding consonant is /t/ and the final vowel is /o/.

Consonant information is included because in Samoan, reanalysis was found to be sensitive to OCP-place effects, which are conditioned by the preceding consonant. I therefore also test for these effects in Māori. As a preview, I find that in Māori, OCP-place effects don't have strong predictive power. Later on in §5.5, I speculate on causes of this discrepancy.

5.4.2   Implementing a phonotactic markedness bias

To implement markedness bias, I follow the same procedure adopted in Chapter 4, which is schematized again in Fig. 5.13. In summary, a phonotactic grammar is trained on monomorphemic roots using the UCLA Phonotactic Learner (UCLAPL; Hayes & Wilson 2008). With the resulting grammar, candidates in the model of reanalysis (e.g. [inumia], [inua], etc.) are assigned harmony scores. These harmony scores then become the constraint violations for a constraint USEPHONOTACTICS, which is biased to have high weight. The input to the phonotactic model was a set of Proto-Polynesian (PPn) roots
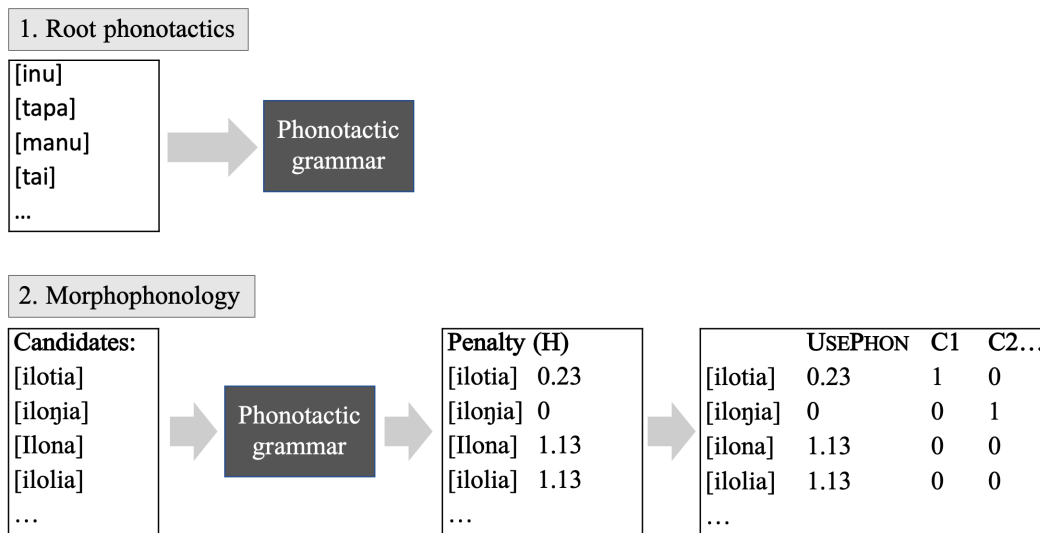
Figure 5.13: Incorporating phonotactic markedness into morphophonological grammar

taken from POLLEX (Greenhill & Clark 2011), which were then modified to reflect the sound changes that took place between PPn and Māori. Just as for Samoan (§4.4.2), each diphthong is represented as a single phoneme. There is also a feature [long], which separates short vowels ([-long]) from long vowels and diphthongs ([+long]).

I trained three phonotactic models:

1. NATURAL CLASS MODEL: The model was limited to learning 50 constraints and given no prespecified constraints. In addition, the model was given a consonant projection (which includes all [-syllabic] segments).

2. HIATUS MODEL: This model was given five prespecified constraints aimed at capturing both the dispreference for heavy vowels and hiatus avoidance. It was then allowed to learn 45 more constraints. The hiatus-related constraints are listed below in (57). Note that the inputs to the phonotactic model are not coded for syllable structure, so the constraint NoOnset cannot be directly implemented. Instead, the constraint *[+syll][+syll] is included to account for word-medial hiatus, and *[+word_boundary][+syll] penalizes initial onsetless syllables.

161

3. HIATUS + OCP MODEL: This model was given only pre-specified constraints, consisting of the same OCP-place constraints used in the Samoan model, along with the five hiatus-related constraints used in the HIATUS model. OCP-place constraints were included because they are known to be present as a statistical tendency in PPn, and because they were active in the reanalysis of Samoan thematic consonants.

(57) *Prespecified constraints in HIATUS model* (segments targeted by each constraint are shown in boldface)

Constraint	Example penalized outputs
*[+long]	**haː**, p**eː**, h**ai**
*[+syllabic][+syllabic]	p**ia**, t**ie**
*[+word_boundary][+syllabic]	**a**ke, **ei**wa, **e**a
*[+long][+syllabic]	h**iːa**, h**eia**ː
*[+syllabic][+long]	h**iaː**, h**aiaː**

5.4.3 Model specifications and results

All model results were averaged over 30 trials, and each model was run for 20 iterations. σ was set to 1 for all constraints.

Just as in Chapter 4, the constraint USEPHONOTACTICS is given a bias towards high weight ($\mu = 3$) in the markedness-biased models; $\mu = 0$ for all other constraints. These models are each compared against a BASELINE model where all constraints are given a μ value of 0. Each baseline model has the same constraint set as its corresponding markedness-biased model (NATURAL CLASS, HIATUS, and HIATUS + OCP). Because all three baseline models behaved similarly, I show just the baseline model corresponding to the HIATUS + OCP model.

Table 5.8 compares the log-likelihood of each model. The rightmost column (ΔL) shows the change in log-likelihood of each model compared to the baseline. Overall, all three markedness-biased models outperform the BASELINE model.

	L	ΔL
BASELINE	-1882.60	–
NATURAL CLASS	-1815.21	67.39
HIATUS	-1702.12	180.48
HIATUS + OCP	-1695.499	187.10

Table 5.8: Model results: log likelihood

The NATURAL CLASS model does only slightly better than the baseline model. On the other hand, both of the models which include hiatus constraints do much better than the BASELINE model. This comparison suggests that the improvement in model fit is mainly driven by the hypothesized markedness constraints, *LONGNUC and NOONSET.

Additionally, the HIATUS + OCP model actually performs slightly better than the HIATUS model. This suggests that while OCP-place constraints do not play a strong role in Māori reanalysis, the directions of reanalysis are consistent with OCP-place constraints.

The two HIATUS models differ from the BASELINE primarily in their prediction for [a]-final stems. This is illustrated in Fig. 5.14, which compares predictions of the BASELINE and HIATUS + OCP models for [a]-final stems. For ease of interpretation, I show predicted probabilities averaged over all preceding consonant conditions. In this figure, [iha-ia] therefore represents all [a]-final stems followed by /-ia/.

Additionally, this figure shows only predictions for /-ia/ and /-tia/, since these are the main candidates of interest. We see that between POc and Māori, there is a drop in the proportion of stems which take /-ia/ and conversely, an increase in stems which take /-tia/. Because the POc inputs show a strong preference for /-ia/, the frequency-matching model is not able to predict this shift towards /-tia/. On the other hand, the HIATUS + OCP model is able to predict change in the right direction.

Note that while the HIATUS + OCP model does trend in the right direction, it does not match the magnitude of change observed in Māori. Future work should consider how this greater magnitude of change should be modeled. For example, in modern Māori, the preference for /-tia/ is much stronger in longer words as opposed to bimoraic words. A

163
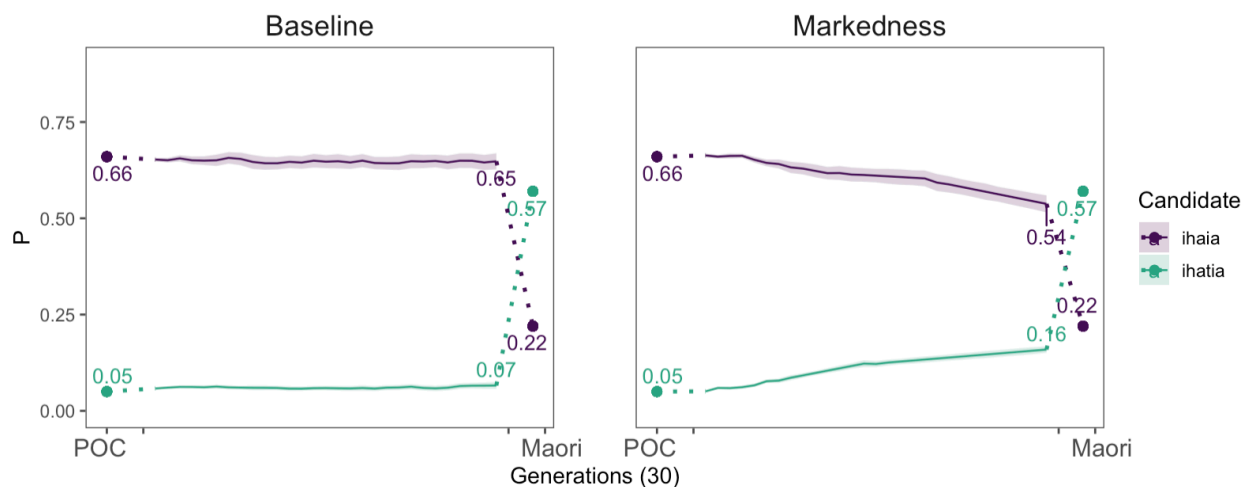
Figure 5.14: Predicted reanalysis in [a]-final stems

model which accounts for prosodic shape of inputs may be able to do better at predicting the observed patterns of reanalysis.

5.5   Comparison of Samoan and Māori

Samoan and Māori CIA suffix allomorphy (and more generally thematic consonant alternations), presumably started out with the same (or at least, very similar) distributional patterns. However, the two languages have settled on different phonological systems. In Samoan, the most frequent (and productive) allomorphs are /-a/ and /-ina/, while in Māori, the most frequent allomorphs are /-a/ and /-tia/. Here, I speculate on possible reasons for this divergence; fully understanding the reasons behind this divergence is beyond the scope of the current project and should be explored in future work.

We can first consider the markedness of different 'default' output forms, as summarized in (58). For the [a]-final stems, stems which take /-ina/ and /-tia/ each violate some relevant markedness constraint, respectively *LongNuc and NoOnset. On the other hand, /-ia/ violates both constraints. From a markedness perspective, /-ia/ is the most marked, while /-a/ and /-ina/ are more closely matched. As such, reanalysis in

164

the direction of either /-ia/→/-ina/ or /-ia/→/-tia/ would be markedness-reducing. In fact, both Samoan and Māori have undergone reanalysis away from /-ia/. In the case of Samoan, /-ia/ is expected to be productive based on the historical development of the CIA suffix. However, it is rarely observed in modern-day Samoan. For Māori, I also show that reanalysis has primarily been away from /-ia/ (and towards /-tia/).

(58)   *Markedness of different CIA allomorphs*

STEM	SUFFIXED	NoOnset	*LongNuc	OCP-cor-son
paka	pa.tai.a	*	*	
	pa.tai.na		*	
	patati.a	*		*

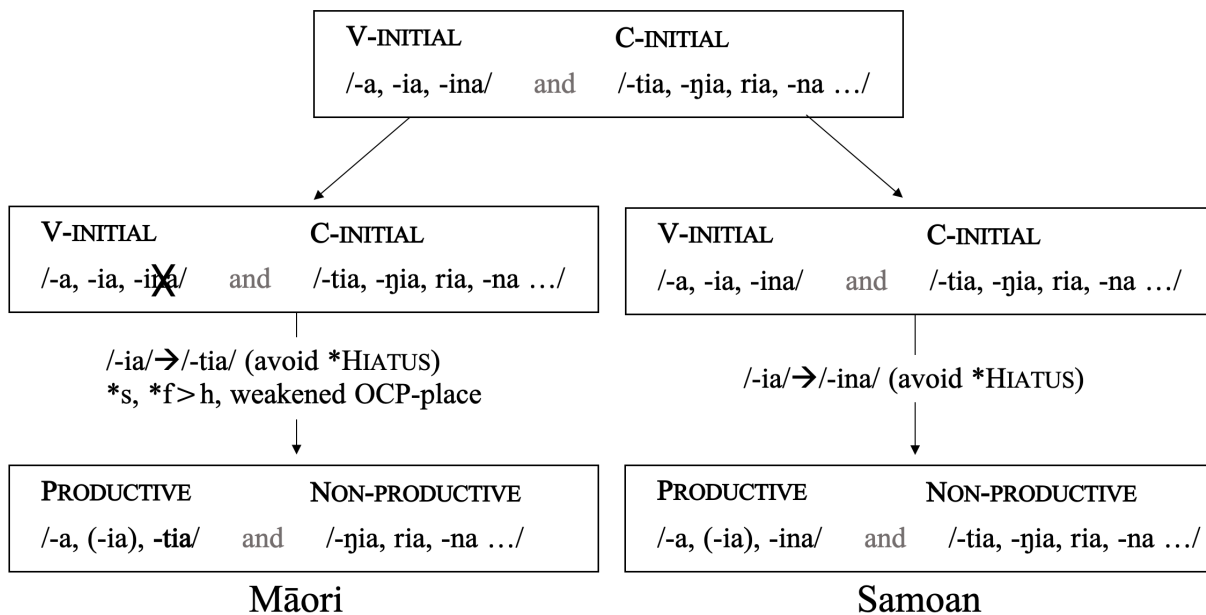Figure 5.15: Hypothesized divergence of CIA allomorphy in Samoan and Māori

While Samoan and Māori have both undergone reanalysis away from /-ia/, they diverge in terms of which allomorph they reanalyzed towards. Fig. 5.15 summarizes a tentative proposal for how this might have happened. First, recall that the CIA allomorph developed from two suffixes, respectively *-i 'TRANSITIVE' and *-a/na '3P.CLITIC'

165

(3rd person pronominal clitic). When the pronominal clitic ceased to be productive, these two suffixes were reanalyzed as a single suffix *ia (and sometimes *ina). It is possible that in Samoan, both *i-a and *i-na were passed down as /-ia/ and /-ina/, while /-ina/ was not productively extended in Māori. As a result, Samoan started with two 'default' vowel-initial suffixes *ia and *ina, while Māori did not. As a result, /-ia/ was reanalyzed to /-ina/ in Samoan, and to /-tia/ in Māori.

Additionally, Samoan and Māori underwent different regular sound changes, which in turn affected the stem phonotactics which served as the basis of reanalysis. In particular, OCP effects are likely to be stronger for coronal obstruents in Samoan than in Māori, where the merger of *s, *f>h would have obscured evidence for OCP effects. This is demonstrated in Table 5.9, which compares a subset of the constraint weights learned by the UCLA Phonotactic Learner for Samoan and Māori. The input to both grammars was a corpus of PPn roots, which were modified to reflect the respective sound changes in Samoan and Māori.

Crucially, while the two phonotactic grammars learn similar weights for the most part, they assign very different weights to *[+cor,-son][+cor,-son], which penalizes OCP-place in coronal obstruents. This constraint has a relatively high weight in Samoan, but no weight at all in Māori. Because OCP effects are weaker for coronal obstruents in Māori, /-tia/ may have had a less restricted distribution. This would have allowed /-tia/ to generalize more easily, to a broader range of environments.

Constraint	Segments targeted	Samoan	Māori
*[+cor,+nasal][+cor,+nasal]	*n...n	0.65	0.47
*[+cor,-son] [+cor,-son]	***{t,s}...{t,s}**	**1.06**	**0**
*[+cor,+son][+cor,+son]	*{n,l}...{n,l}	0.97	1.04
*[+cor,-cont][+cor,-cont]	*{t,n}...{t,n}	0.12	0.37

Table 5.9: Weights learned by UCLA Phonotactic Learner (Input = PPn roots, reflecting respective sound changes in Samoan and Māori)

In sum, although Māori and Samoan started with what was historically a very similar system of CIA-allomorphy, small divergences in regular sound change may have resulted

in the two different systems we observe today. Although it is not the focus of the current project, models of reanalysis could also be helpful for testing hypotheses about how historically similar morphophonological systems diverge.

CHAPTER 6

Conclusion

6.1 Summary of results

In this dissertation, I investigated the effects of markedness bias (i.e. constraints on output forms) on the reanalysis of morphophonological paradigms. Existing models of reanalysis are frequency-matching; that is, they predict that reanalysis should occur in a way that matches probabilistic distributions within a paradigm. I propose that in fact, reanalysis responds to (at least) two factors: both frequency-matching and the reduction of markedness.

Chapter 2 outlines a model of reanalysis which is used in subsequent chapters to quantitatively test hypotheses about how reanalyses has occurred in different languages. This model uses MaxEnt, a probabilistic implementation of Optimality Theory (Smolensky 1986; Goldwater & Johnson 2003). Bias is implemented as a Gaussian prior following Wilson (2006) and White (2013, 2017). More concretely, I vary μ for constraints, and give the relevant markedness constraints a higher μ than competing faithfulness constraints. The model also has an iterative component to simulate the cumulative effect of reanalysis over generations of speakers.

I additionally make the distinction between 'universal' markedness and 'active' markedness, or markedness effects already active in language-specific phonotactics. I propose that only active markedness can affect reanalysis. All the case studies presented in this dissertation are consistent with the principle of active markedness.

This dissertation reports three case studies (Chapter 3-5) where reanalysis is argued

to be modulated by the effects of a markedness bias. The first case study, Malagasy weak stems (Chapter 3), shows a clear case of reanalysis towards the statistically dispreferred alternant, which I argue to be motivated by avoidance of intervocalic stops. The results of this chapter are consistent with the active markedness proposal, as intervocalic stops are also dispreferred in Malagasy stem phonotactics.

The second and third case studies concern reanalysis of thematic consonant alternations in Samoan and Māori. Allomorphy involving thematic consonants is more complex than the Malagasy example. While this makes trends in reanalysis more difficult to observe, it also allows for testing of more nuanced hypotheses about which markedness generalizations speakers can extract from stem phonotactics.

In Samoan (Chapter 4), reanalysis is generally towards the vowel-initial allomorphs, as predicted by frequency-matching models. However, reanalysis is also modulated by markedness effects, such that suffixed forms which violate OCP-place are more likely to be reanalyzed.

I also test several phonotactic grammars and find that grammars which are restricted to learning OCP constraints over natural classes outperform less restrictive alternatives. This result suggests that speakers are not just picking up on any statistical regularity in the stem to inform the direction of reanalysis. Instead, they may be biased towards picking up patterns that are typologically motivated or rooted in phonetic naturalness.

In Māori (Chapter 5), I find that reanalysis is towards the consonant-initial /-tia/ allomorph, rather than the vowel-initial allomorphs (which are the expected targets of reanalysis based on historical distributions). I argue that this change is driven by avoidance of long syllables in hiatus. Interestingly, Samoan and Māori started out with the same system of CIA allomorphy, but diverged in how reanalysis took place. In Section 5.5, I discuss some possible reasons for this, such as differences in stem phonotactics that emerged as a result of regular sound changes.

6.2 Markedness effects as synchrony vs. diachrony

The case studies discussed in this dissertation provide novel evidence for markedness bias in phonological learning. In my modeling implementation, I assume that markedness bias is present in the synchronic grammar; reanalysis arises because at each generation of speakers, bias in the learner's synchronic grammar causes mislearning of paradigms. Over time, generations of incremental change result in the restructuring of a pattern.

However, as Glewwe (2019) points out, these types of markedness effects are hard to find in experiments, where the evidence for markedness in learning is mixed. This has led some people to argue that there is no synchronic bias for less marked outputs, and that shared cross-linguistic tendencies in avoiding marked structures are only the result of sound change (Ohala 1993; Hale & Reiss 2000; Blevins 2004).

Another possibility is that markedness effects do exist in the synchronic grammar, but are of such a small magnitude that they cannot be reliably found in an experimental setting. Instead, it takes more robust data, such as findings from change over time, to observe markedness bias in phonology.

Additionally, my results suggest that markedness effects in morphophonology may be stronger when they are supported by stem phonotactics. Put another way, future research on markedness bias should make a distinction between markedness effects that have support from stem phonotactics, and ones that don't. There is some recent experimental work supporting this distinction. For example, in a recent AGL study, Chong (2021) find evidence that speakers can extend static phonotactic generalizations to alternation patterns.

6.3 Future directions

In principle, there are various ways in which markedness and stem phonotactics could influence morphophonology. On one hand, there is the question whether a markedness

constraint needs to be active or not in the phonotactics to influence reanalysis; throughout this dissertation, I have characterized this distinction as one of 'active' vs. 'inactive' markedness. On the other hand, when speakers draw on stem phonotactics to aid in the learning of alternations, are they extracting principles rooted in phonetic naturalness, or any sort of statistical generalization present in stems? These two parameters ('active' vs. 'inactive' and 'natural vs. 'unnatural') are summarized in (59) below.

(59) *Typology of possible reanalyses*

	Active	Inactive
Natural	✓	?
Unnatural	?	?

As summarized in (59), my results are consistent with the idea that reanalyses are constrained by both parameters, in that markedness effects must both be active in phonotactics and phonetically natural. First, in all three case studies, the markedness effects present in reanalysis are also active in the stem phonotactics, supporting the active markedness restriction. Additionally, in my model of Samoan reanalysis, an OCP phonotactic grammar outperformed other less restrictive phonotactic grammars. This suggests that speakers are restricted in which phonotactic principles they can extract and make use of. Future work should expand on the typology of markedness effects in reanalysis, to confirm whether both restrictions hold true crosslinguistically. Experimental work could also be done to supplement findings from reanalyses over time, to see if similar restrictions are present in learners' synchronic grammars.

In general, if an active markedness restriction is present in reanalysis, then other aspects of morphophonology could be similarly constrained by phonotactics. One potential area to examine is apparent cases of The Emergence of the Unmarked (TETU; McCarthy & Prince 1994), where some marked structure is generally allowed in a language, but

banned in particular contexts; in other words, the effects of a markedness constraint are obscured in some contexts, but "emerge" in other contexts. For example, a language may allow codas within stems, but disallow then in reduplicants. In their discussion of TETU effects, McCarthy & Prince (1994) characterize the relevant markedness constraints as universal (i.e. part of a universal constraint set present in all grammars). If future work finds a correlation between stem phonotactics and TETU effects, this could challenge the universal markedness approach.

My model of reanalysis also predicts that the certainty of a generalization interacts with the strength of markedness effects. In particular, markedness effects should be stronger when there is more uncertainty in a paradigm. This prediction is hard to confirm using data on language change, but could potentially be teased apart in experimental settings.

Finally, this dissertation briefly touches on how a model of reanalysis can be used to not just predict the direction of reanalysis, but also model the divergence of morphophonological systems over time. In particular, Māori and Samoan started with the same input distributions, but ended up with different synchronic patterns of allomorphy. These results could be compared against reanalysis in other Polynesian languages. For example, Māori and Samoan remain relatively conservative in maintaining thematic consonants. In Hawaiian, on the other hand, the CIA suffix has leveled to just one predictable suffix (/ʔia/). This crosslinguistic variation makes the CIA suffix allomorphy a valuable case study for testing predictions about how markedness bias can affect reanalysis.

Bibliography

Adelaar, Alexander. 1989. Malay influence on Malagasy: Linguistic and culture-historical implications. *Oceanic Linguistics, A Special Issue on Western Austronesian Languages* 28(1). 1–46.

Adelaar, Alexander. 2009. Malagasy. In Martin Haspelmath & Uri Tadmor (eds.), *World loanword database (WOLD)*, Max Planck digital library.

Adelaar, Alexander. 2012. Malagasy phonological history and Bantu influence. *Oceanic Linguistics* 51(1). 123–159.

Adelaar, Alexander. 2013. Malagasy dialect divisions: Genetic versus emblematic criteria. *Oceanic Linguistics* 52(2). 457–480.

Adelaar, Alexander K. 1994. Malay and Javanese loanwords in Malagasy, Tagalog and Siraya (Formosa). *Bijdragen tot de taal-, land-en volkenkunde* 150(1). 50–66.

Albright, Adam. 2002a. A restricted model of UR discovery: Evidence from Lakhota. *Ms, University of California at Santa Cruz* .

Albright, Adam. 2005. The morphological basis of paradigm leveling. In Laura Downing, Tracy Alan Hall & Renate Raffelsiefen (eds.), *Paradigms in phonological theory*, 17–43. Oxford University Press.

Albright, Adam. 2008. Explaining universal tendencies and language particulars in analogical change. In Jeff Good (ed.), *Linguistic universals and language change*, 144–184. Oxford University Press.

Albright, Adam. 2010. Base-driven leveling in Yiddish verb paradigms. *NLLT* 28(3). 475–537.

Albright, Adam & Bruce Hayes. 2003. Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90(2). 119–161.

Albright, Adam C. 2002b. *The identification of bases in morphological paradigms*. PhD dissertation, University of California, Los Angeles.

Albro, Daniel Matthew. 2005. *Studies in computational Optimality Theory, with special reference to the phonological system of Malagasy*. PhD dissertation, University of California, Los Angeles.

Alderete, John & Mark Bradshaw. 2013. Samoan root phonotactics: Digging deeper into the data. *Linguistic Discovery* 11(1).

Apoussidou, Diana. 2006. *The learnability of metrical phonology*. PhD dissertation, University of Amsterdam.

Baayen, R Harald, Richard Piepenbrock & Leon Gulikers. 1996. *The CELEX lexical database (cd-rom)*. University of Pennsylvania.

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

Baker, Adam. 2008. Computational approaches to the study of language change. *Language and Linguistics Compass* 2(2). 289–307.

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

Bauer, Winifred. 1993. *Maori*. Routledge.

Bauer, Winifred. 1997. *The Reed reference grammar of Maori*. Reed Books.

Becker, Michael. 2009. *Phonological trends in the lexicon: the role of constraints*. PhD dissertation, University of Massachusetts, Amherst.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. *Language* 84–125.

Benua, Laura. 1995. Identity effects in morphological truncation. In Suzanne Urbanczyk Jill N. Beckman & Suzanne Urbanczyk (eds.), *University of massachusetts occasional papers 18: Papers in optimality theory*, 77–136. Amherst: GLSA.

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

Berkley, Deborah M. 1994. The OCP and gradient data. *Studies in the Linguistic Sciences* 1/2. 59–72.

Berkley, Deborah M. 2000a. *Gradient ocp effects*. PhD dissertation, Northwestern University.

Berkley, Deborah Milam. 2000b. *Gradient obligatory contour principle effects*. PhD dissertation, Northwestern University.

Biggs, Bruce. 1961. The structure of New Zealand Maaori. *Anthropological Linguistics* 1–54.

Biggs, Bruce. 1989. Towards a study of Māori dialects. In R Harlow & R Hooper (eds.), *Vical 1 oceanic languages: Papers from the fifth international conference on austronesian linguistics*, Linguistic Society of New Zealand.

Biggs, Bruce. 2013. *The Complete English–Maori Dictionary, 4th edition*. Auckland University Press.

Blevins, Juliette. 1994. A phonological and morphological reanalysis of the Maori passive. *Te Reo* 37. 29–53.

Blevins, Juliette. 1995. The syllable in phonological theory. In John A. Goldsmith (ed.), *The handbook of phonological theory*, 245–306. Blackwell.

Blevins, Juliette. 2004. *Evolutionary phonology: The emergence of sound patterns*. Cambridge University Press.

Blevins, Juliette. 2008. Consonant epenthesis: natural and unnatural histories. In Jeff Good (ed.), *Language universals and language change*, 79–107. Oxford University Press.

Blust, Robert & Stephen Trussel. 2010. Austronesian comparative dictionary, web edition. *Blust's Austronesian Comparative Dictionary Website* .

Blythe, Richard A & William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language* 269–304.

Boersma, Paul. 1998. *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.

Boersma, Paul. 2007. Some listener-oriented accounts of h-aspiré in french. *Lingua* 117(12). 1989–2054.

Boersma, Paul & Silke Hamann. 2009. Cue constraints and their interactions in phonological perception and production. *Phonology in perception* 15. 55–110.

Bolognesi, Roberto. 1998. *The Phonology of Campidanian Sardinian: A Unitary Account of a Self-organizing Structure* HIL dissertations. Holland Academic Graphics. `https://books.google.com/books?id=\_ModAQAAIAAJ`.

Boyce, Mary. 2006. *A corpus of modern spoken Māori*. PhD dissertation, Victoria University of Wellington.

Brighton, Henry. 2002. Compositional syntax from cultural transmission. *Artificial life* 8(1). 25–54.

Buckley, Eugene. 1997. Tigrinya root consonants and the OCP. *University of Pennsylvania Working Papers in Linguistics* 4(3). 3.

Bybee, J. 2003. *Phonology and language use* Cambridge Studies in Linguistics. Cambridge University Press.

Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10. 425–455.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge University Press.

Bybee, Joan L & Carol Lynn Moder. 1983. Morphological classes as natural categories. *Language* 59(2). 251–270.

Bybee, Joan L & Dan I Slobin. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58(2). 265–289.

Calderone, Basilio, Nabil Hathout & Olivier Bonami. 2021. Not quite there yet: Combining analogical patterns and encoder-decoder networks for cognitively plausible inflection. *arXiv:2108.03968* .

Casali, Roderic F. 1997. Vowel elision in hiatus contexts: Which vowel goes? *Language* 73(3). 493–533.

Cheung, Hintat. 2000. Three to four-years old children's perception and production of Mandarin consonants. *Language and Linguistics* 1(2). 19–38.

Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. ERIC.

Chong, Adam J. 2019. Exceptionality and derived environment effects: a comparison of Korean and Turkish. *Phonology* 36(4). 543–572.

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

Chung, Sandra. 1978. *Case marking and grammatical relations in Polynesian*. PhD dissertation, Harvard University.

Churchward, Spencer. 1951. *A Samoan grammar*. Spectator Publishing Company.

Clahsen, Harald, Fraibet Aveledo & Iggy Roca. 2002. The development of regular and irregular verb inflection in Spanish child language. *Journal of child language* 29(3). 591–622.

Clapp, Edward Bull. 1906. On correption in hiatus. *Classical Philology* 1(3). 239–252.

Clark, David Ross. 1973. *Aspects of proto-Polynesian syntax*. PhD dissertation, University of California, San Diego.

Coetzee, Andries W & Joe Pater. 2008a. Lexically ranked OCP-Place constraints in Muna. ROA-842.

Coetzee, Andries W & Joe Pater. 2008b. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *NLLT* 26. 289–337.

Colantoni, Laura & Jeffrey Steele. 2005. Liquid asymmetries in French and Spanish. *Toronto Working Papers in Linguistics* 24.

Coleman, John & Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *3rd meeting of the ACL special interest group in computational phonology: Proceedings of the workshop*, 49–56. Association for Computational Linguistics.

Cook, Kenneth William. 1988. *A cognitive analysis of grammatical relations, case, and transitivity in Samoan*. University of California, San Diego.

Cutler, Anne, Andrea Weber, Roel Smits & Nicole Cooper. 2004. Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America* 116(6). 3668–3678.

Daelemans, Walter, Jakub Zavrel, Kurt Van Der Sloot & Antal Van den Bosch. 2004. Timbl: Tilburg memory-based learner. *Tilburg University* .

Dahl, Otto Christian. 1951. *Malgache et Maanjan: une comparaison linguistique*. Egede Instituttet.

Daugherty, Kim G & Mark S Seidenberg. 1994. Beyond rules and exceptions. In Susan D Lima, Roberta Corrigan & Gregory K Iverson (eds.), *The reality of linguistic rules*, 353–388. John Benjamins Publishing.

Dempwolff, Otto. 1929. *Das austronesische Sprachgut in den polynesischen Sprachen*. G. Kolff.

Denison, David. 2003. Log(ist)ic and simplistic s-curves. In R Hickey (ed.), *Motives for language change*, Cambridge University Press.

Dresher, B Elan. 1980. The Mercian Second Fronting: a case of rule loss in Old English. *Linguistic Inquiry* 11(1). 47–73.

Dresher, B Elan. 1985. *Old English and the theory of phonology*, vol. 4. New York: Garland.

Dresher, B Elan. 2015. Rule-based generative historical phonology. *The Oxford handbook of historical phonology* 501–521.

Dyen, Isidore. 1951. Proto-Malayo-Polynesian *Z. *Language* 27(4). 534–540.

Dziwirek, Katarzyna. 1989. Malagasy phonology and morphology. *Linguistic Notes from La Jolla* 15. 1–30.

Eberhard, David M, Gary F Simons & Charles D. Fennig (eds). 2023. *Ethnologue: Languages of the world (26th edition)*. Dallas, Texas: SIL International. `http://www.ethnologue.com`.

Eddington, D. 2004. *Spanish phonology and morphology: Experimental and quantitative perspectives*. John Benjamins Publishing Company.

Eddington, David. 1996. Diphthongization in Spanish derivational morphology: An empirical investigation. *Hispanic Linguistics* 8(1). 1–13.

Eddington, David. 1998. Spanish diphthongization as a non-derivational phenomenon. *Rivista di Linguistica* 10. 335–354.

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.

Ernestus, Mirjam Theresia Constantia & R Harald Baayen. 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language* 79(1). 5–38.

Erwin, Sean. 1996. Quantity and moras: An amicable separation. *UCLA Occasional Papers in Linguistics* 17. 2–30.

Evans, Bethwyn. 2001. *A study of valency-changing devices in Proto-Oceanic*. PhD dissertation, Research School of Pacific and Asian Studies, Australian National University.

Flemming, Edward. 2004. Contrast and perceptual distinctiveness. In Bruce Hayes, Donca Steriade & Robert Kirchner (eds.), *Phonetically based phonology*, 232–276. Cambridge University Press.

Frisch, Stefan A, Janet B Pierrehumbert & Michael B Broe. 2004. Similarity avoidance and the OCP. *Language & Linguistic Theory* 22(1). 179–228.

Frisch, Stefan A & Bushra Adnan Zawaydeh. 2001. The psychological reality of OCP-Place in Arabic. *Language* 77(1). 91–106.

Fylstra, Daniel, Leon Lasdon, John Watson & Allan Waren. 1998. Design and use of the Microsoft Excel Solver. *Interfaces* 28(5). 29–55. doi:10.1287/inte.28.5.29.

Gallagher, Gillian & Jessica Coon. 2009. Distinguishing total and partial identity: Evidence from Chol. *NLLT* 27. 545–582.

Gallagher, Gillian, Maria Gouskova & Gladys Camacho Rios. 2019. Phonotactic restrictions and morphology in Aymara. *Glossa* 4(1).

Garrett, Andrew. 2008. Paradigmatic uniformity and markedness. In Andrew Garrett (ed.), *Linguistic universals and language change*, 125–143. Oxford University Press.

Glewwe, Eleanor Rachel. 2019. *Bias in phonotactic learning: Experimental studies of phonotactic implicationals*. University of California, Los Angeles.

Gnanadesikan, Amalia. 2004. Markedness and faithfulness constraints in child phonology. In Rene Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 73–108. Cambridge University Press.

Goldwater, Sharon & Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the stockholm workshop on variation within optimality theory*, 111–120.

Gouskova, Maria. 2018. Morphology and phonotactics. In *Oxford research encyclopedia of linguistics*, Oxford University Press. doi:10.1093/acrefore/9780199384655.013.613.

Gouskova, Maria & Michael Becker. 2013. Nonce words show that Russian yer alternations are governed by the grammar. *NLLT* 31. 735–765.

Green, Roger. 1966. Linguistic subgrouping within Polynesia: The implications for prehistoric settlement. *The Journal of the Polynesian Society* 75(1). 6–38.

Greenberg, Joseph. 1950. The patterning of root morphemes in Semitic. *Word* 6. 162–181.

Greenberg, Joseph. 1966. *Universals of language*. MIT Press.

Greenhill, Simon J & Ross Clark. 2011. POLLEX-online: The Polynesian lexicon project online. *Oceanic Linguistics* 551–559.

Griffiths, Thomas L & Michael L Kalish. 2007. A Bayesian view of language evolution by iterated learning. *Cognitive Science* 31. 441–480.

Gunkel, Dieter C & Kevin Ryan. 2011. Hiatus avoidance and metrification in the Rigveda. *Proceedings of the 22nd Annual UCLA Indo-European Conference* .

Hale, Kenneth. 1968. Review of Hohepa (1967)—'a profile generative grammar of Maori'. *Journal of the Polynesian Society* 77. 83–99.

Hale, Kenneth. 1973. Deep-surface canonical disparities in relation to analysis and change: An Australian example. In T Sebeok (ed.), *Current trends in linguistics 11*, The Hague: Mouton.

Hale, Mark & Charles Reiss. 2000. "Substance abuse" and "dysfunctionalism": Current trends in phonology. *Linguistic Inquiry* 31(1). 157–169.

Halle, Morris. 1959. *The sound pattern of Russian: A linguistic and acoustical investigation.* Mouton.

Hanson, Kristin. 2001. Quantitative meter in English: the lesson of Sir Philip Sidney. *English Language & Linguistics* 5(1). 41–91.

Hansson, Gunnar Ólafur. 2007. On the evolution of consonant harmony: The case of secondary articulation agreement. *Phonology* 24(1). 77–120.

Hare, Mary & Jeffrey L. Elman. 1995. Learning and morphological change. *Cognition* 56(1). 61–98.

Harlow, Ray. 2007. *Maori: A linguistic introduction.* Cambridge University Press.

Hayes, Bruce. 1995. *Metrical stress theory: principles and case studies.* University of Chicago Press.

Hayes, Bruce. 2004. Phonological acquisition in optimality theory: the early stages. In Rene Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 158–203. Cambridge University Press.

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

Hayes, Bruce. 2022. Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics* 8. 473–494.

Hayes, Bruce & Jinyoung Jo. 2020. Balinese stem phonotactics and the subregularity hypothesis. Ms, UCLA.

Hayes, Bruce & Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. *Phonology* 23(1). 59–104.

Hayes, Bruce, Péter Siptár, Kie Zuraw & Zsuzsa Londe. 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language* 822–863.

Hayes, Bruce & James White. 2015. Saltation and the P-map. *Phonology* 32(2). 267–302.

Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39(3). 379–440.

Hayes, Bruce, Colin Wilson & Anne Shisko. 2012. Maxent grammars for the metrics of Shakespeare and Milton. *Language* 88(4). 691–731.

Hellberg, Staffan. 1978. Unnatural phonology. *Journal of Linguistics* 14(2). 157–177.

Hock, Hans Henrich. 1991. *Principles of historical linguistics*. De Gruyter Mouton.

Hohepa, Patrick Wahanga. 1967. *A profile generative grammar of Maori*. Indiana University Press.

Hovdhaugen, Even et al. 1986. The chronology of three Samoan sound changes. In *Focal ii: Papers from the fourth international conference on austronesian linguistics*, Pacific Linguistics.

Howe, Penelope. 2021. Central Malagasy. *Journal of the International Phonetic Association* 51(1). 103–136.

Hudson, Alfred B. 1967. The Barito isolects of Borneo; a classification based on comparative reconstruction and lexicostatistics. *Southeast Asia Program (Dept. of Far Eastern Studies), Data Paper no. 68* .

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

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

Inkelas, Sharon, Orhan Orgun & Cheryl Zoll. 1997. The implications of lexical exceptions for the nature of grammar. In Iggy Roca (ed.), *Derivations and constraints in phonology*, 393–418. Oxford, Clarendon.

Ito, Chiyuki & Naomi H. Feldman. 2022. Iterated learning models of language change: A case study of sino-korean accent. *Cognitive Science* 46(4). e13115.

Ito, Junko & Armin Mester. 2003. On the sources of opacity in OT: coda processes in German. *The syllable in optimality theory* 271–303.

Jackson, Dan & Garrison W Cottrell. 1997. Attention and U-shaped learning in the acquisition of the past tense. In *Proceedings of the nineteenth cognitive science conference*, 325–30.

Jakobson, Roman. 1963. Essais de linguistique générale. *Les Etudes Philosophiques* 18(4).

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

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

Johnson, Mark. 2002. Optimality-theoretic lexical functional grammar. In Paola Merlo & Suzanne Stevenson (eds.), *The lexical basis of sentence processing: Formal, computational and experimental issues*, John Benjamins Pub. Co.

Jones, Carwyn. 2012. Ko Aotearoa tēnei: A report into claims concerning New Zealand law and policy affecting Māori culture and identity. *Victoria University of Wellington Legal Research Paper No. 69* .

Jones, Ōiwi Parker. 2008. Phonotactic probability and the maori passive: A computational approach. In *Proceedings of the tenth meeting of ACL special interest group on computational morphology and phonology*, 39–48.

Jun, Jongho & Jeehyun Lee. 2007. Multiple stem-final variants in Korean native nouns and loanwords. *Journal of the Linguistic Society of Korea* 47. 159–187.

Jun, Sun-Ah. 1994. The status of the lenis stop voicing rule in Korean. *Theoretical issues in Korean linguistics* 101–114.

Kabak, Barış & Irene Vogel. 2001. The phonological word and stress assignment in Turkish. *Phonology* 18(3). 315–360.

Kager, René. 1996. On affix allomorphy and syllable counting. In Kleinhenz Ursula (ed.), *Interfaces in phonology*, 155–171. Akademie Verlag.

Kager, René. 2000. Stem stress and peak correspondence in Dutch. *Optimality Theory: phonology, syntax, and acquisition* 121–150.

Kang, Yoonjung. 2006. Neutralizations and variations in Korean verbal paradigms. *Harvard Studies in Korean Linguistics* 11. 183–196.

Kaplan, Abby. 2010. *Phonology shaped by phonetics: The case of intervocalic lenition*. PhD dissertation, University of California, Santa Cruz.

Katz, Jonah. 2016. Lenition, perception and neutralisation. *Phonology* 33(1). 43–85.

Kawahara, Shigeto. 2012. Lyman's Law is active in loanwords and nonce words: Evidence from naturalness judgment studies. *Lingua* 122(11). 1193–1206.

Kawahara, Shigeto, Hajime Ono & Kiyoshi Sudo. 2006. Consonant co-occurrence restrictions in Yamato Japanese. *Japanese/Korean Linguistics* 14. 27–38.

Kean, Mary-Louise. 1975. *The theory of markedness in generative grammar*. PhD dissertation, Massachusetts Institute of Technology.

Keenan, Edward L & Cecile Manorohanta. 2001. A quantitative study of voice in Malagasy. *Oceanic linguistics* 67–84.

Keenan, Edward L & Maria Polinsky. 2017. Malagasy (austronesian). *The handbook of morphology* 563–623.

Kenstowicz, Michael. 1996. Base-identity and uniform exponence: alternatives to cyclicity. In Jacques Durand & Bernard Laks (eds.), *Current trends in phonology: Models and methods*, 363–394. Salford: University of Salford.

Kenstowicz, Michael. 1997. Uniform exponence: Extension and exemplification. In *Selected papers from the Hopkins optimality workshop 1997, University of Maryland working papers in linguistics*, vol. 5, 139–154.

Kenstowicz, Michael & Charles Kisseberth. 1977. *Topics in phonological theory*. Elsevier.

Kibre, Nicholas. 1998. Formal property inheritance and consonant/zero alternations in Maori verbs. *Ms., Rutgers Optimality Archive 285* .

Kim, Jong-Kyoo. 2000. *Quantity-sensitivity and feature-sensitivity of vowels: A constraint-based approach to Korean vowel phonology*. PhD dissertation, Indiana University.

Kiparsky, Paul. 1965. *Phonological change*. PhD dissertation, Massachusetts Institute of Technology.

Kiparsky, Paul. 1978. Analogical change as a problem for linguistic theory. *Studies in the Linguistic Sciences Urbana, Ill* 8(2). 77–96.

Kiparsky, Paul. 1982. Lexical morphology and phonology. In I.-S. Yang (ed.), *Linguistics in the morning calm*, 3–91. Seoul: Hansin.

Kiparsky, Paul. 1988. Phonological change. In Frederick J. Newmeyer (ed.), *Linguistics: The cambridge survey: Volume 1, linguistic theory: Foundations*, CUP Archive.

Kiparsky, Paul. 1989. Sprung rhythm. In Paul Kiparsky & Gilbert Youmans (eds.), *Rhythm and meter*, 305–340. Elsevier.

Kiparsky, Paul. 1997. Covert generalization. In *Mediterranean morphology meetings*, vol. 1, 65–76.

Kiparsky, Paul. 2012. Grammaticalization as optimization. *Grammatical change: Origins, nature, outcomes* 15–51.

Kirby, Simon. 2001. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. *IEEE Transactions on Evolutionary Computation* 5(2). 102–110.

Kirchner, Robert M. 1998. *An effort-based approach to consonant lenition*. PhD dissertation, UCLA.

Kisseberth, Charles W. 1970. On the functional unity of phonological rules. *Linguistic Inquiry* 1(3). 291–306.

Kroch, Anthony S. 1989. Reflexes of grammar in patterns of language change. *Language variation and change* 1(3). 199–244.

Krupa, Victor. 1967. On phonemic structure of morpheme in Samoan and Tongan. *Beitrèage zur Linguistik und Informationsverarbeitung* 12. 72–83.

Krupa, Viktor. 1966. The phonemic structure of bi-vocalic morphemic forms in Oceanic languages. *The Journal of the Polynesian Society* 75(4). 458–497.

Krupa, Viktor. 1971. The phonotactic structure of the morph in Polynesian languages. *Language* 47(3). 668–684.

Kuo, Jennifer. 2023. Evidence for prosodic correspondence in the vowel alternations of tgdaya seediq. *Phonological Data and Analysis* 5(3). 1–31.

Kuryłowicz, Jerzy. 1945. La nature des procès dits «analogiques». *Acta linguistica* 5(1). 15–37.

Kuryłowicz, Jerzy & Margaret Winters. 1947. The nature of the so-called analogical processes. *Diachronica* 12(1). 113–145.

de La Beaujardière, Jean-Marie. 2004. Malagasy dictionary and encyclopedia of Madagascar.

Labov, William. 1994. *Principles of linguistic change, volume 1: Internal factors*. Wiley-Blackwell.

de Lacy, Paul. 2002. Maximal words and the Maori passive. In Norvin Richards (ed.), *Proceedings of the austronesian formal linguistics association (afla) 8*, MIT Working Papers in Linguistics.

de Lacy, Paul. 2003. Maximal words and the Maori passive. In John McCarthy (ed.), *Optimality theory in phonology: A reader*, 495–512. Blackwell.

Lahiri, Aditi & B Elan Dresher. 1999. Open syllable lengthening in West Germanic. *Language* 678–719.

Lavoie, Lisa M. 2001. *Consonant strength: Phonological patterns and phonetic manifestations*. Routledge.

Levelt, Clara C, Niels O Schiller & Willem J Levelt. 2000. The acquisition of syllable types. *Language acquisition* 8(3). 237–264.

Levelt, Clara C & Ruben Van de Vijver. 1998/2004. Syllable types in cross-linguistic and developmental grammars. In René Kager, Joe Pater & Wim Zonneveld (eds.), *Constraints in phonological acquisition*, 204–218. Cambridge: Cambridge University Press.

Lewis, Paul M., Gary F. Simons & Charles D. Fennig. 2014. *Ethnologue: Languages of Asia*. SIL international.

Lichtenberk, Frantisek. 2001. On the morphological status of thematic consonants in two Oceanic languages. In Joel Bradshaw & Kenneth L Rehg (eds.), *Issues in Austronesian morphology: A festschrift for Byron W. Bender*, 123–147. Pacific Linguistics.

Ling, Charles & Marin Marinov. 1993. Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. *Cognition* 49(3). 235–290.

MacWhinney, Brian & Jared Leinbach. 1991. Implementations are not conceptualizations: Revising the verb learning model. *Cognition* 40(1-2). 121–157.

Mahdi, Waruno. 1988. *Morphophonologische besonderheiten und historische phonologie des malagasy*, vol. 20. D. Reimer.

Malouf, Robert. 2002. A comparison of algorithms for maximum entropy parameter estimation. *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)* .

Mańczak, Witold. 1957. Tendances générales des changements analogiques. *Lingua* 7. 298–325.

Mańczak, Witold. 1980. Laws of analogy. In J Disiak (ed.), *Historical morphology*, 283–288. The Hague: Mouton.

Marck, Jeff. 1996. Eastern Polynesian subgrouping today. In Janet Davidson, Geoffrey Irwin, Foss Leach, Andrew Pawley & Dorothy Brown (eds.), *Oceanic culture history: Essays in honour of Roger Green*, 491–511. New Zealand Journal of Archaeology.

Marck, Jeff. 1999. Revising Polynesian linguistic subgrouping and its culture history implications. In Roger Blench & Matthew Spriggs (eds.), *Archaeology and language IV: Language change and cultural transformation*, 95–122. Routledge.

Marck, Jeff. 2000. *Topics in Polynesian language and culture history*. Pacific Linguistics.

Marcus, Gary F, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu & Harald Clahsen. 1992. Overregularization in language acquisition. *Monographs of the society for research in child development* 1–178.

Martin, Andrew. 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. *Language* 87(4). 751–770. doi:10.1353/lan.2011.0096.

Mayer, Connor, Adeline Tan & Kie Zuraw. 2022. *maxent.ot: A package for doing Maximum Entropy Optimality Theory in R (Version 0.1.0)*. `https://doi.org/10.5281/zenodo.7246367`. Computer software.

McCarthy, John. 1981. The role of the evaluation metric in the acquisition of morphology. In C. L. Baker & J McCarthy (eds.), *The logical problem of language acquisition*, 218–248. MIT Press.

McCarthy, John J. 1988. Feature geometry and dependency: A review. *Phonetica* 45(2-4). 84–108.

McCarthy, John J. 1994. The phonetics and phonology of Semitic pharyngeals. In Patricia Keating (ed.), *Phonological structure and phonetic form*, 191–233. Cambridge University Press.

McCarthy, John J & Alan Prince. 1993. *Generalized alignment*. Springer.

McCarthy, John J & Alan Prince. 1994. The emergence of the unmarked: Optimality in prosodic morphology. In *Proceedings of the north east linguistics society 24*, 333–379.

McCarthy, John J. & Alan S. Prince. 1995. Faithfulness and Reduplicative Identity. In Laura Walsh Dickey Jill N. Beckman & Suzanne Urbanczyk (eds.), *Papers in optimality theory*, 249–384. Amherst: GLSA.

Merchant, Nazarré Nathaniel. 2008. *Discovering underlying forms: Contrast pairs and ranking*. Rutgers State University of New Jersey.

Mester, Armin R. 1986. *Studies in tier structure*. PhD dissertation, University of Massachusetts Amherst.

Mester, R Armin. 1994. The quantitative trochee in Latin. *NLLT* 12(1). 1–61.

Milner, George Bertram. 1966. *Samoan Dictionary; Samoan-English, English-Samoan*. ERIC.

Minkova, Donka. 1982. The environment for open syllable lengthening in Middle English. *Folia Linguistica Historica* 16(Historica-vol-3-1). 29–58.

Minkova, Donka & Robert Stockwell. 2003. English vowel shifts and 'optimal' diphthongs: Is there a logical link? *Optimality Theory and language change* 169–190.

Moder, Carol Lynn. 1992. *Productivity and categorization in morphological classes*. PhD dissertation, State University of New York at Buffalo.

Moore-Cantwell, Claire. 2008. Samoan thematic consonants and the -Cia suffix.

Moore-Cantwell, Claire & Joe Pater. 2016. Gradient exceptionality in Maximum Entropy Grammar with lexically specific constraints. *Catalan Journal of Linguistics* 15. 53–66.

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

Moreton, Elliott & Joe Pater. 2012a. Structure and substance in artificial-phonology learning, part I: Structure. *Language and linguistics compass* 6(11). 686–701.

Moreton, Elliott & Joe Pater. 2012b. Structure and substance in artificial-phonology learning, part {II}: Substance. *Language and linguistics compass* 6(11). 702–718.

Mosel, Ulrike & Even Hovdhaugen. 1992. *Samoan reference grammar*. Scandinavian Univ. Press.

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

Ngata, Hori M. 1971. *English-Maori Dictionary*. Wellington, N.Z.: Learning Media. `https://www.teaching.co.nz/page/mta-nz-about-ngata-dictionary`.

Nigg, Benno M. 1994. General comments about modelling. In Benno M. Nigg & Walter Herzog (eds.), *Biomechanics of the musculo-skeletal system*, 367–379. John Wiley & Sons.

Niyogi, Partha. 2006. *The computational nature of language learning and evolution*. MIT press Cambridge, MA.

Nosofsky, Robert M. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical psychology* 34(4). 393–418.

Nosofsky, Robert M. 2011. The generalized context model: An exemplar model of classification. In Emmanuel M. Pothos & Andy J. Wills (eds.), *Formal approaches in categorization*, 18–39. Cambridge University Press.

Odden, David. 2007. The unnatural tonology of Zina Kotoko. In Tomas Riad & Carlos Gussenhoven (eds.), *Tones and tunes, vol. 1: Typological studies in word and sentence prosody*, 63–89. Berlin: Mouton de Gruyt.

Oh, Y, S Todd, C Beckner, J Hay, J King & J Needle. 2020. Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific reports* 10(1). 1–9.

Ohala, John J. 1993. Sound change as nature's speech perception experiment. *Speech Communication* 13(1-2). 155–161.

O'Neill, Timothy. 2015. *The phonology of Betsimisaraka Malagasy*. PhD dissertation, University of Delaware.

Padgett, Jaye. 1991a. *Stricture in feature geometry*. PhD dissertation, University of Massachusetts, Amherst.

Padgett, Jaye. 1991b. *Stricture in feature geometry*. Dissertations in Linguistics. CSLI Publications.

Parker, George Williams. 1883. *A concise grammar of the Malagasy language.* Trubner.

Paster, Mary. 2005. Subcategorization vs. output optimization in syllable-counting allomorphy. In *Proceedings of the 24th west coast conference on formal linguistics,* vol. 24, 326. Cascadilla Proceedings.

Paster, Mary. 2009. Explaining phonological conditions on affixation: Evidence from suppletive allomorphy and affix ordering. *Word structure* 2(1). 18–37.

Pater, Joe. 2007. The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In Leah Bateman & Adam Werle (eds.), *UMOP: Papers in optimality theory iii,* 1–36. University of Massachusetts.

Pater, Joe. 2008. Morpheme-specific phonology: Constraint indexation and inconsistency resolution. In Steve Parker (ed.), *Phonological argumentation: Essays on evidence and motivation,* 1–33. Equinox.

Pater, Joe, Robert Staubs, Karen Jesney & Brian Smith. 2012. Learning probabilities over underlying representations. In *Proceedings of the twelfth meeting of the special interest group on computational morphology and phonology (SIGMORPHON2012),* Association for Computational Linguistics.

Pater, Joe & Anne-Michelle Tessier. 2005. Phonotactics and alternations: Testing the connection with artificial language learning. *University of Massachusetts Occasional Papers in Linguisitcs* 31. 1–16.

Pawley, Andrew. 1962. The person-markers in Samoan. *Te Reo* 5. 52–56.

Pawley, Andrew. 1966. Polynesian languages: A subgrouping based on shared innovations in morphology. *The Journal of the Polynesian Society* 75(1). 39–64.

Pawley, Andrew. 1967. The relationships of Polynesian Outlier languages. *The Journal of the Polynesian Society* 76(3). 259–296.

Pawley, Andrew. 2001. Proto polynesian *-CIA. In Joel Bradshaw & Kenneth L Rehg (eds.), *Issues in Austronesian morphology: A festschrift for Byron W. Bender*, Pacific Linguistics.

Petersen, Stacy. 2016. Vowel dispersion in English diphthongs: Evidence from adult production. In *Proceedings of the annual meetings on phonology*, vol. 3, .

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

Pierrehumbert, Janet. 2002. Word-specific phonetics. In C Gussenhoven & N Warner (eds.), *Laboratory phonology VII*, 101—140. Berlin: Mouton de Gruyter.

Pierrehumbert, Janet B. 2006. The statistical basis of an unnatural alternation. *Laboratory phonology* 8. 81–107.

Polinsky, Maria & Ezra van Everbroeck. 2003. Development of gender classifications: Modeling the historical change from Latin to French. *Language* 79(2). 356–390.

Prasada, Sandeep & Steven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and cognitive processes* 8(1). 1–56.

Pratt, George. 1862/1893. *A Samoan dictionary: English and Samoan, and Samoan and English, with a short grammar of the Samoan dialect*. London Missionary Society's Press.

Prince, Alan & Paul Smolensky. 1993. *Optimality theory: Constraint interaction in generative grammar*. Wiley Online Library.

Rasin, Ezer & Roni Katzir. 2016. On evaluation metrics in optimality theory. *Linguistic Inquiry* 47(2). 235–282.

Rasin, Ezer & Roni Katzir. 2020. A conditional learnability argument for constraints on underlying representations. *Journal of Linguistics* 56(4). 745–773.

Rasoloson, Janie & Carl Rubino. 2005. Malagasy. In K Alexander Adelaar & Nikolaus Himmelmann (eds.), *The Austronesian languages of Asia and Madagascar,* 456–488. Routledge Abingdon.

Roberts-Kohno, Rosalind Ruth. 2000. *Kikamba phonology and morphology.* PhD dissertation, The Ohio State University.

Rumelhart, David E & James L McClelland. 1987. Learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In B MacWhinney (ed.), *Mechanisms of language aquisition,* Lawrence Erlbaum Associates, Inc.

Russell, Kevin. 1995a. Morphemes and candidates in Optimality Theory. *Ms., Rutgers Optimality Archive* 44.

Russell, Kevin. 1995b. Morphemes and candidates in Optimality Theory. *Rutgers Optimality Archive* 44.

Ryan, Peter M. 2012. *The raupō dictionary of modern Māori.* Penguin Random House New Zealand Limited.

Sanders, Gerald. 1990. On the analysis and implications of Maori verb alternations. *Lingua* 80(2). 149–196. doi:10.1016/0024-3841(90)90019-H.

Sanders, Gerald. 1991. Levelling and reanalysis in the history of Polynesian passive formations. *Journal of the Polynesian Society* 100. 71–90.

Schumacher, R Alexander & Janet B Pierrehumbert. 2021. Familiarity, consistency, and systematizing in morphology. *Cognition* 212. 104512.

Seidl, Amanda & Eugene Buckley. 2005. On the learning of arbitrary phonological rules. *Language learning and development* 1(3-4). 289–316.

Selkirk, Elisabeth. 1980. Prosodic domains in phonology: Sanskrit revisited. In Mark Arnoff & Mary-Louise Kean (eds.), *Juncture,* 107–129. Anma Libri.

Sihler, Andrew L. 1995. *New comparative grammar of Greek and Latin*. Oxford University Press.

Simões, Maria Cecília Perroni & Carol Stoel-Gammon. 1979. The acquisition of inflections in Portuguese: A study of the development of person markers on verbs. *Journal of child language* 6(1). 53–67.

Siptár, Péter. 2003. Hungarian Yod. *Acta Linguistica Hungarica* 50(3-4). 457–473.

Siptár, Péter & Miklós Törkenczy. 2000. *The phonology of Hungarian*. Oxford University Press.

Skousen, R. 1989. *Analogical modeling of language*. Springer Netherlands.

Smith, Brian W. 2015. *Phonologically conditioned allomorphy and UR constraints*. PhD dissertation, University of Massachusetts Amherst.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory. Tech. rep. Colorado Univ at Boulder Dept of Computer Science.

Smolensky, Paul & Géraldine Legendre. 2006. *The harmonic mind: From neural computation to optimality-theoretic grammar (cognitive architecture), vol. 1*. MIT press.

Stanley, Richard. 1967. Redundancy rules in phonology. *Language* 393–436.

Stats NZ. 2018. 2018 census totals by topic – national highlights (updated). Retrieved from `www.stats.govt.nz` (May 14, 2023).

Steriade, Donca. 1997. Lexical conservatism. *Linguistics in the morning calm* 157–179.

Steriade, Donca. 2001. The phonology of perceptibility effects: The p-map and its consequences for constraint organization.

Steriade, Donca. 2009. The phonology of perceptibility effects: the P-map and its consequences for constraint organization. In Kristin Hanson & Sharon Inkelas (eds.), *The

nature of the word: Studies in honor of paul kiparsky, 151–180. Cambridge, MA: MIT Press.

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

Suzuki, Keiichiro. 1998. *A typological investigation of dissimilation*. PhD dissertation, The University of Arizona.

Tabor, Whitney. 1994. *Syntactic innovation: A connectionist model*. PhD dissertation, Stanford University.

Tan, Adeline. 2022. Concurrent hidden structure & grammar learning. *Proceedings of the Society for Computation in Linguistics* 5(1). 55–64.

Tesar, Bruce, John Alderete, Graham Horwood, Nazarré Merchant, Koichi Nishitani & Alan S Prince. 2003. Surgery in language learning. In *Proceedings of the twenty-second west coast conference on formal linguistics*, 477–490. Cascadilla Press.

Tesar, Bruce & Alan Prince. 2003. Using phonotactics to learn phonological alternations. *CLS* 39(2). 241–269.

Tsao, Feng-Ming, Ching-Yun Lee, Yi-Hsin Hsieh & Chin-Yeh Chiu. 2009. Assessing stop and lexical tone perception in preschool children and relationship with word development. *Journal of the Speech-Language-Hearing Association of Taiwan* 24. 39–57. doi: doi:10.6143/JSLHAT.2009.12.03.

Vennemann, Theo. 1972. Rule inversion. *Lingua* 29(3-4). 209–42.

Violette, Pere L. 1880. *Dictionnaire samoa-français-anglais*. Maisonneuve.

Wang, Marilyn D. & Robert C. Bilger. 1973. Consonant confusions in noise: A study of perceptual features. *JASA* 54(5). 1248–1266.

Warner, Natasha, James M McQueen & Anne Cutler. 2014. Tracking perception of the sounds of English. *JASA* 135(5). 2995–3006.

Weinreich, Uriel, William Labov & Marvin Herzog. 1968. *Empirical foundations for a theory of language change*. University of Texas Press.

Wheeler, Max W. 2005. *The phonology of Catalan*. Oxford University Press.

White, James. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language* 93(1). 1–36.

White, James C. 2013. *Bias in phonological learning: Evidence from saltation*. PhD dissertation, UCLA.

Williams, Herbert W. 1957. *A dictionary of the Maori language, 6th edition*. Government Printer. `https://nzetc.victoria.ac.nz/tm/scholarly/tei-WillDict.html`.

Williams, Herbert W. 1971. *A dictionary of the Maori language, 7th edition*. Wellington, N.Z. : Government Printer.

Williams, William. 1844. *A dictionary of the New Zealand language, 1st edition*. C.M. Society.

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

Wilson, Colin & Marieke Obdeyn. 2009. Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms, Johns Hopkins University.

Wilson, William H. 1985. Evidence for an outlier source for the Proto Eastern Polynesian pronominal system. *Oceanic Linguistics* 24(1/2). 85–133.

Wilson, William H. 2012. Whence the East Polynesians? Further linguistic evidence for a Northern Outlier source. *Oceanic Linguistics* 51(2). 289–359.

Wolf, Matthew Adam. 2008. *Optimal interleaving: Serial phonology-morphology interaction in a constraint-based model*. PhD dissertation, University of Massachusetts Amherst.

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

Yang, Charles D. 2000. Internal and external forces in language change. *Language variation and change* 12(3). 231–250.

Yip, Moira. 1989. Feature geometry and co-occurrence restrictions. *Phonology* 6. 349–374.

Zamuner, Tania S. 2006. Sensitivity to word-final phonotactics in 9-to 16-month-old infants. *Infancy* 10(1). 77–95.

Zipf, George Kingsley. 1999. *The psycho-biology of language: An introduction to dynamic philology*. Routledge.

Zuraw, Kie. 2002. Aggressive reduplication. *Phonology* 19(3). 395–439.

Zuraw, Kie. 2003. Probability in language change. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 139–176. MIT Press.

Zuraw, Kie. 2010a. A model of lexical variation and the grammar with application to tagalog nasal substitution. *NLLT* 28(2). 417–472. doi:10.1007/s11049-010-9095-z.

Zuraw, Kie. 2010b. A model of lexical variation and the grammar with application to Tagalog nasal substitution. *NLLT* 28(2). 417–472.

Zuraw, Kie. 2013. *map constraints. Ms, University of California, Los Angeles.

Zuraw, Kie & Bruce Hayes. 2017. Intersecting constraint families: an argument for harmonic grammar. *Language* 93(3). 497–548.

Zuraw, Kie, M Yu Kristine & Robyn Orfitelli. 2014. The word-level prosody of Samoan. *Phonology* 31(2). 271–327.

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

Zymet, Jesse. 2020. Malagasy OCP targets a single affix: Implications for morphosyntactic generalization in learning. *Linguistic Inquiry* 51(3). 624–634.