Phonological reanalysis is guided is guided by markedness: the case of Malagasy weak stems

Abstract. A key goal in phonology is to understand the factors that affect phonological learning. This paper addresses the issue by examining how paradigms are reanalyzed over time. Malagasy has a class of stems, called weak stems, where final consonants alternate when suffixed. Comparison of historical and modern Malagasy shows that weak stem paradigms have undergone extensive reanalysis in a way that cannot be predicted by the probabilistic distribution of alternants. This poses a problem for existing quantitative models of reanalysis, where reanalysis is always towards the most probable alternant. I argue instead that reanalysis in Malagasy is driven by both distributional factors and a markedness bias. To capture the Malagasy pattern, I propose a maximum entropy learning model (Goldwater & Johnson, 2003), with a markedness bias implemented via the model's prior probability distribution. This biased model successfully predicts the direction of reanalysis in Malagasy, outperforming purely distributional models.

1 Introduction

Understanding the extent to which different biases affect phonological acquisition is a central question in phonology. This question has been addressed extensively through experimental work (e.g. Wilson, 2006; Moreton & Pater, 2012b,a) and research on child language acquisition (e.g. Singleton & Newport, 2004; Peperkamp et al., 2006). Since Kiparsky's seminal work on phonological change (1965; 1968; 1978, et seq), it has been recognized that studying language change over time can also give us insight into the factors that drive phonological learning. The data may be harder to interpret due to the large time depth, but also potentially offer more contextual validity than experimental work. Insights from language change can therefore complement experimental and acquisition research.

The current study focuses on a specific type of change, reanalysis in paradigms. Morphological paradigms can have neutralizing alternations that cause ambiguity in one or more slots of the paradigm. For example, Middle High German (MHG) had a well-known process of final obstruent devoicing that created ambiguity in non-suffixed forms (Sapir, 1915, p.237; Kiparsky, 1968, p.177, etc.). As demonstrated by the examples in (1a), given a non-suffixed MHG stem with a final voiceless obstruent, the final obstruent could either surface as voiceless (e.g. zak∼zakə), or show a voicing alternation (e.g. vek∼vegə).

(1) Reanalysis of obstruent voicing in Yiddish (nominative sg. vs pl. paradigm)

(a)			(b)			(c)		
MHG		→	Early Yiddish		→	Modern Yiddish		
sg.	pl.		sg.	pl.		sg.	pl.	
ve**k**	ve**g**ə		ve**k**	ve**g**(ə)		ve**g**	ve**g**ən	'way'
za**k**	za**k**ə		za**k**	ze**k**(ə)		za**k**	ze**k**	'sack'

Neutralizing alternations like this can be challenging to the language-learning child, and be prone to reanalysis over time. This was the case for voicing alternations in Yiddish, a direct descendant of MHG. Final obstruent devoicing was present in early Yiddish (1b), but subsequently lost in Modern Yiddish, where the singular forms were reanalyzed to remove neutralization. As shown in (1c), the voicing value of the plural was reintroduced to the singular (Albright, 2010).

Notably, there are relatively few quantitative models that can make strong, language-specific predictions about the output and direction of reanalysis. Existing models predict reanalysis to be solely based on the probabilistic distribution of segments. In these models, reanalysis is always in the direction of the more probable alternant.

In the current study, however, I find that for Malagasy, the direction of reanalysis contradicts the predictions of purely distributional models. Specifically, in a class of stems called ‘weak stems’, there has been extensive reanalysis in a direction that is not predicted by distributional properties in the lexicon. I argue that reanalysis in Malagasy is sensitive to both distributional and markedness effects. Building on these results, I propose a constraint-based model of reanalysis which has a markedness bias.

The rest of the paper is organized as follows: §2 introduces existing models of reanalysis, and presents the descriptive facts of Malagasy weak stems. In §3, I present results of a corpus study comparing historical Malagasy forms with modern Malagasy data, to show that reanalysis has occurred in a direction that cannot be predicted by purely inductive models of reanalysis. Finally, §4 proposes a model of reanalysis which incorporates a markedness learning bias.

2 Background

2.1 Quantitative approaches to modeling reanalysis

Existing quantitative models of reanalysis (or more generally of morphophonological paradigm learning) are inductive, and therefore predict change to be driven purely by statistical distributions. One representative model of this variety is the Minimal Generalization Learner (MGL; Albright & Hayes, 2002; Albright, 2002; Albright & Hayes, 2003, et seq.).

The MGL first compares different members of the paradigm, and learns word-specific rules mapping from one form to another. With regards to the MHG pattern introduced above, the MGL would generate rules like in (2). When forms share the same change, the model finds what features they share in common, and generalizes rules based on these shared features. For example, a rule ∅→ə/[−voiceless, −continuant]_# may be generated from comparison of forms (a) and (b). The result is a system of stochastic rules which predict the inflected form of a paradigm given an input base.

(2) Word-specific rules learned by the MGL for MHG

	sg.	pl.	word-specific rule
(a)	za**k**	za**k**ə	∅→ə/ vek_#
(b)	mu**t**	za**t**ə	∅→ə/ mut_#
(c)	ve**k**	ve**g**ə	k→gə/ ve_#

In the MGL, reanalysis occurs when the grammar derives the incorrect output for certain derived forms, and these errors come to replace the older, exceptional forms. This model has been shown to explain the direction of historical restructuring in various languages, including Lakhota (Albright, 2008), Yiddish (Albright, 2010), and Korean (Kang, 2006). Details of model implementation can be found in Albright & Hayes (2003). What is important to note is that this model learns rules inductively, and predicts reanalysis to always be in the direction of the statistically most probable outcome, given the distribution of sounds in a paradigm.

Albright’s model is rule-based, and generates sets of rules that predict the outcome of paradigm reanalysis. An alternative analogical approach is exemplified by the Generalized Context Model (GCM Nosofsky, 2011). This approach is ‘similarity-based’, meaning that in principle, any words that are similar enough to each other can serve as the basis for reanalysis. Broadly speaking, similarity-based models are less restrictive than rule-based models, and are potentially able to capture a wider range of effects Albright & Hayes (2003). However, both approaches predict that reanalysis will match the distributions of the input data.

Inductive learning is also possible in stochastic constraint-based models such as Maximum Entropy Harmonic Grammar (MaxEnt; Goldwater & Johnson, 2003; Smolensky, 1986). As a preview, in §4, an inductive constraint-based model will be used as a baseline, and compared to more models which incorporate learning biases.

2.2 Malagasy phonology and weak stem alternations

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). The following study uses data from Official Malagasy (OM), which is the standardized, institutional dialect that is based on the dialect spoken in the capital city Merina. All subsequent descriptions and analysis will assume data from OM.

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.

Malagasy has been studied extensively. The phonetic system is described by Howe (2021), and basic facts on the morphology and phonology are documented in work such as Keenan & Polinsky (2017) for OM, and O'Neill (2015) for 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). Moreover, the history of Malagasy can be traced in some detail through the work of Austronesianists (e.g. Dahl, 1951; Mahdi, 1988; Adelaar, 2013). Additionally, dictionary data is digitized in the Malagasy Dictionary and Encyclopedia of Madagascar (MDEM; de La Beaujardière 2004), which compiles data from multiple Malagasy dictionaries. Historical comparative data is also available the Austronesian Comparative Dictionary (ACD; Blust & Trussel, 2010).

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

2.2.1 Malagasy phonology

Malagasy words have a strict (C)V syllable structure, where codas are not allowed. Word stress is phonemic but generally penultimate, though there are exceptions to 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 because /ua/ and /au/ sequences have merged to /o/ in OM (Howe, 2021).

The consonants of Malagasy are given in Table 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 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

	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 1: Malagasy consonant chart

alveolar trill in most dialects including OM, but is often realized as a tap [ɾ] in casual speech Howe (2021).[1]

2.2.2 Weak stems

Malagasy has a class of forms that Keenan & Polinsky (2017) refer to as weak stems. These roots have antepenultimate stress (if long enough), and always end in one of the three 'weak syllables' ʈʂa, ka, or na.[2]

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 2, using the active and passive forms of verbs. 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 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.

pattern		**active (m + stem)**	**passive (stem + ana)**	
na ∼	n	manˈɖʐavi**n**a	anɖʐaˈvi**n**ana	'to bear leaves'
	m	maˈnanɖʐa**n**a	aˈnɖʐá**m**ana	'to try'
ka ∼	h	maˈngata**k**a	angaˈta**h**ana	’to ask for’
	f	maˈnaha**k**a	anaˈha**f**ana	’to scatter’
ʈʂa ∼	r	miána**ʈʂ**a	ianá**r**ana	'to learn'
	t	maˈnanɖʐa**ʈʂ**a	anaˈnɖʐa**t**ana	'to promote'
	f	maˈnɖʐaku**ʈʂ**a	anɖʐaˈku**f**ana	'to cover'

Table 2: Patterns of consonant alternation in Malagasy weak stems

The standard formal analysis for weak stems is that they are underlyingly consonant-final (Albro, 2005). For example, the surface forms [m-iˈanaʈʂa]∼[iaˈnar-ana] would have the stem UR /ianar/, with surface forms derived as in (3). First, all words are assigned penultimate stress, and

[1] My personal observations in work with a consultant matches Howe's phonetic descriptions.

[2] According to Howe (2021), the final vowel of weak stems is often devoiced or reduced.

the stem-final consonant is neutralized to [ʈ͜ʂ], [k], or [n] (here, /r/ neutralizes to [ʈ͜ʂ]). In the suffixed form, /r/ is medial and therefore protected from neutralization. Finally, an epenthetic /a/ is added to resolve the violation against codas (counterbleeding the final-C neutralization). Antepenultimate stress falls out naturally from the rule ordering, where stress assignment precedes vowel epenthesis. As I discuss below, the analysis in (3) is in fact a recapitulation of the historical development of weak stem alternations.

(3) Derivation for surface forms of /ianar/ in a formal analysis of weak stems

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

2.2.3 Historical development of weak stem alternations

The linguistic history of Malagasy has been studied in detail. The following description summarizes findings from a large body of scholarship, 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 Kalimantan into the Comoro Islands. Contact with Bantu during this migration significantly influenced Malagasy morphology, and is largely thought to have caused the development of final open syllables in Malagasy. For most final consonants, epenthesis of a final vowel removed final codas, resulting in the weak stems of current 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 (4) trace a subset of weak stems through these stages, to illustrate the historical development of some weak stem alternations.

(4a) illustrates the development of a ʈ͜ʂa∼t alternating weak stems, which historically end in voiceless coronal stops, in this case *t. Final **-t* neutralized to **-ʈ͜ʂ* 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 ʈ͜ʂa∼t alternation.

In (4b), on the other hand, the PMP stem ends in **D* [d̪]. In the non-suffixed form, this final consonant devoiced to **-t*, and then neutralized to *ʈ͜ʂ*. In the suffixed form, **D* lenited to *r* due to regular sound change (*D>r; Adelaar 2012). This was followed by final vowel epenthesis, resulting in the observed ʈ͜ʂa∼r alternation. 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 (4c-4d) 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 (4c). The development of ka∼f alternating follows from a similar process, given in (4d). First, **-p* and **-k* neutralized to *-k* word-finally. This was followed by spirantization from **p>f*.

(4) Examples: historical basis of final consonant alternations; changes relevant to the consonant alternation are given in parentheses.[3]

a. ʈʂa∼t alternation[4]

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’

c. 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’

d. 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	< *j [g^j],*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: Weak stem alternants and corresponding historical consonants

Table 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.

[3] 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.

[4] 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].

As a caveat, most consonant-final PMP forms reflect as weak stems in Malagasy, but there are three 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 have ∅∼C alternations in modern Malagasy (e.g. [ˈlalu∼laˈlu**v**-ana] < *lalaw, 'pass without stopping'). 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 4.

Coda resolved by...	**PMP cons.**	**Mlg alternation**	**Example**
Vowel epenthesis	*-k,*-g	ka∼h	baˈliaka∼ibaliˈah-ana
	*-p, *-b	ka/ʈ͜ʂa∼f	ˈhirika∼ hiˈrif-ana
	-t,-c	ʈ͜ʂa∼t	ˈharatra∼ haˈrat-ana
	*-d, *-D,*-j	ʈ͜ʂa∼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 phoneme)	∅∼s	miˈlefa∼leˈfa**s**-ana

Table 4: Malagasy reflexes of stem-final PMP consonants

3 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 Malagasy. In the following section, I discuss the predicted outcome of reanalysis under a distributional 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 ʈ͜ʂa 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 ʈ͜ʂa∼t alternating stem is reanalyzed as r-alternating. Conversely, reanalysis could happen in the opposite direction, where a historically ʈ͜ʂa∼r alternating stem becomes t-alternating. (5) summarizes the possible outcomes of reanalysis, given the hypothetical ʈ͜ʂa-final weak stem [ˈpakuʈ͜ʂa].

(5) *Possible directions of reanalysis for ʈ͜ʂa-final weak stems (example stem: [ˈpakuʈ͜ʂa])*

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 ʈʂ̥a, 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 -ʈʂ̥a 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.

Results are based off of comparison of historical and modern Malagasy data. Historical data is taken from the Austronesian Comparative Dictionary (ACD; Blust & Trussel, 2010) and Adelaar (2012). 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.[5]

Section 3.1 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 Section 3.2. Section 3.3 provides additional indirect evidence on the directions of reanalysis using data from modern Malagasy.

3.1 Predicted reanalyses under an inductive approach

In a purely inductive model of morphophonological learning, reanalysis would always 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 5 shows the distribution of 215 PMP protoforms with final consonants that correspond to Malagasy weak stems. Protoforms with no known Malagasy reflexes were excluded. Results are organized by which alternant each PMP final consonant would correspond.

There is one complication when [f] is the alternant. Historically, stem-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 ʈʂ̥a-final weak stems. In Table 5, all PMP forms ending in labial stops are assumed to correspond to ka-final weak stems in Malagasy. This simplification should not impact the analysis, since ʈʂ̥a∼f alternating forms make up a very small proportion of ʈʂ̥a-final weak stems.

From this data, we see that ka-final weak stems have many more h-alternating forms, na-final weak have more non-alternating forms, and ʈʂ̥a-final weak stems have more t-alternating forms. An inductive approach predicts that reanalysis should be in the direction of these more frequent

[5] The primary dictionaries that the MDEM sources from were all published from 1885-1998; more details can be found in https://en.mondemalgache.org/bins/sources.

Type	alternant	count	%	Predicted reanalysis
ka	h (<*k)	50	89.3%	f→h
	f (<*p,*b)	6	10.7%	
na	m (<*m)	8	8.4%	m→n
	n (<*n,*ŋ)	87	91.6%	
ʈʂa	r (<*j,*r,*d,*ɖ)	17	26.6%	r→t
	t (<*t)	47	73.4%	

Table 5: Expected distribution of Malagasy weak stem alternants, based on the distribution of PMP final consonants.

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 5.

Mahdi's (1988) findings on dissimilatory effects in weak stems are also partially replicated in the PMP data. Consider (6), which tabulates 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. In all eight protoforms coded as containing a preceding [r], the expected Malagasy alternant is [t]. Of the forms where the expected alternant is [r], none were coded as containing a preceding [r].

(6)

	does stem have [r] (<*r,*d,*ɖ,*j)?	
alternant	yes	no
t	8	39
r	0	17

For ka-final weak stems, however, there is not a clear dissimilatory pattern in PMP. 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]. However, when the expected alternant is [h], around 10% (n = 5/50) of protoforms have a preceding *k. When the expected alternant is [f], a similar proportion of forms (n = 1/9,11%) have a preceding *k.

(7)

	does stem have h (<*k)?	
alternant	yes	no
f	1	8
h	5	45

3.2 Observed directions of reanalysis

In this section, PMP stems are compared to their weak stem reflexes, and cases of mismatch are used to infer the direction of reanalyses. 25 (out of 205) protoforms were excluded because their Malagasy reflexes are not weak stems with known/productive suffixed forms. Of these excluded forms, the majority (n = 20) were nasal-final, and correspond to na-final weak stems. The data here is also supplemented with 54 early Malay and Javanese loanwords from the World Loanword Database (WOLD; Adelaar, 2009) and Adelaar (1994). These loans were introduced to Malagasy before the migration to the Comoro Islands (and therefore before the development of weak stems). Tables 6-8 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 6 shows the results for na-final weak stems. The top two rows show cases where the observed alternant in Malagasy matches the historical final consonant, indicating that reanalysis has not taken place. The bottom two rows show cases of mismatch, where reanalysis has occurred. There are relatively few renalyses, but most (3/4, 75%) are in the direction of m→n (e.g. [ˈizi**na**∼iˈzi**n**ina] <*qie**m** 'shade, darkness'). This is in line with the predictions of an inductive approach. There is one exception, the stem [ˈteno**na**∼teˈno**m**-ina] (<*tenun) 'to weave/be woven'. Given the lack of data, it is hard to tell what the cause is.[6]

		alternant			
	match?	expected	observed	change	count
(a)	yes	n	n		60
(b)	yes	m	m		6
(c)	no	n	m	n→m	1
(d)	no	m	n	m→n	3

Table 6: Expected vs. observed alternant of na-final stems, based on known protoforms/loanwords

Table 7 shows the reanalyses for ka-final weak stems. The rightmost column 'has h?' gives the number of forms in each category where final [ka] is immediately preceded by an [h], with only one intervening vowel. Once again, there are relatively few cases of reanalyses. However, as seen in row (c), most reanalyses (4/5, 80%) are in the direction of f→h (e.g. [ˈheni**k**a∼heˈni**h**ina] <*gene**p** 'to be satisfied with'), in line with predictions of an inductive approach. Of all protoforms expected to show f-alternation, 3/8 (38%) have undergone reanalysis of f→h.

From the data, it is unclear whether a dissimilatory effect is active in reanalysis. The 5 f-alternating forms that were not reanalyzed did not have [h] as the consonant nearest to the alternant, so h-dissimilation cannot be the factor blocking them from reanalysis. However, the one case of reanalysis in the direction of h→f could potentially be attributed to h-dissimilation. This word, [ˈlauka∼laˈufana] (<*lahuk) 'meat/relish eaten with rice', historically had a preceding [h] which was subsequently elided in PSEB. This pattern, if present, is surprising because it is not supported by distributional information in PMP.

		alternant				
	match?	expected	observed	change	count	has h?
(a)	yes	f	f		5	0
(b)	yes	h	h		50	5
(c)	no	f	h	f→h	4	1
(d)	no	h	f	h→f	1	1

Table 7: Expected vs. observed alternant of ka-final stems, based on known protoforms/loanwords

Table 8 shows results for ʈʂa-final weak stems. The rightmost column, 'has r?', indicates whether there is an [r] in the stem. For ʈʂa-final stems, extensive reanalysis has occurred. Overwhelmingly, [r] appears to be the preferred alternant. In fact, as seen in row (e) , 30 forms

[6] 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)

underwent renalysis in the direction of t→r (e.g. [ˈaka**ʈ͜ʂ**a∼aˈka**r**ina] <*aŋka**t**, ‘ascent’). This accounts for over half (n = 30/53, 57%) of originally t-alternating stems. Moreover, as seen in row (f), when ʈ͜ʂa∼f alternating forms are reanalyzed, it is always in the direction of f→r (e.g. [ˈtaka**ʈ͜ʂ**a∼taˈka**r**ina] < *taqka**p**, ‘attain, seize’).

		alternant				
	match?	expected	observed	change	count	has r?
(a)	yes	r	r		18	0
(b)	yes	t	t		23	14 (61%)
(c)	yes	f	f		6	1
(d)	no	r	t	r→t	1	1
(e)	no	t	r	t→r	30	0
(f)	no	f	r	f→r	3	0

Table 8: Expected vs. observed alternant of ʈ͜ʂa-final stems, based on known proto-forms/loanwords

Additionally, r-dissimilation appears to be active in the reanalysis of ʈ͜ʂa-final weak stems, in that reanalysis to [r] is blocked if the stem has a preceding [r]. As seen in row (e) and (f), 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 (row (b)), a relatively large proportion (n = 14/23, 61%) had a preceding [r] (e.g.[ˈsu**r̠**i**ʈ͜ʂ**a∼suˈ**r̠**i**t**ana] <*curit, ‘mark’).

The only example of reanalysis in the direction of r→t (row (d)) is likely motivated by r-dissimilation. The reanalyzed form [ˈsanɖ͜ʐa**ʈ͜ʂ**a∼anaˈnɖ͜ʐa**t**ana] (<sanda**r**, Malay loan) does not have an [r] in modern Malagasy, but [nɖ͜ʐ] sequences are historically [nr], and only affricated to [nɖ͜ʐ] in a later stage of PSEB (Proto Southeast-Barito).

The direction of reanalysis in ʈ͜ʂa-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.

3.3 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.

Table 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 is consistent with the finding that reanalyses have been in the direction of m→n, increasing the relative frequency of non-alternating na-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. Although no such effect was found in PMP (or in the attested reanalyses), h-dissimilation does seem to be present in modern Malagasy. This is illustrated in Fig. 1, which shows the distribution of alternants for ka-final stems, by whether or not the consonant nearest to the alternant is [h]. When there is an immediately preceding [h], the observed alternant is always [f]. In contrast, when the

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

Table 9: Proportion of alternants for modern Malagasy weak stems

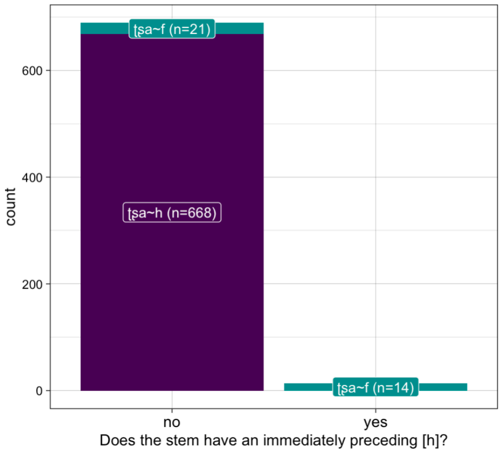

Figure 1: Distribution of alternants in ka-final weak stems

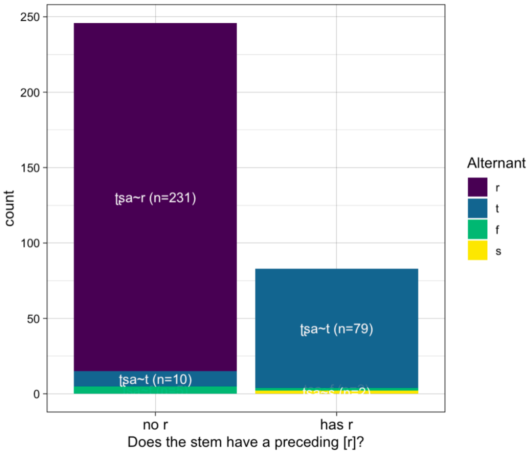

Figure 2: Distribution of alternants in ʈʂa-final weak stems

stem does not have a preceding [h], only 3% (n = 21/689) stems have [f] as the alternant. Based on these results, dissimilation could have affected reanalyses of ka-final stems.

The data in Table 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. 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 dissimilatory pattern. Specifically, final ʈʂa *never* alternates with [r] if there is already an [r] in the stem. In contrast, when there is no preceding [r], there is a strong, near-exceptionless preference for alternation with [r]. Overall, the distribution of alternants in modern Malagasy supports the finding that reanalysis in ʈʂa-final weak stems is in the direction of t→r, except when blocked by r-dissimilation.

3.4 Markedness effects on reanalysis of ʈʂa stems

For the ʈʂa-final weak stems, reanalysis in the direction of t→r cannot be explained by an inductive 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 historical 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 at 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 explained as the result of 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).

One alternative possibility is that speakers are driven by a perceptual bias, rather than a markedness bias (Steriade, 2009 [2001]; Wilson, 2006; White, 2013). That is, if the retroflex affricate [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]

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).

3.5 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 (8). On one hand, reanalysis of na- and ka-final weak stems is largely predictable from distributional probabilities.

[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).

(8) Summary: directions of reanalysis in Malagasy

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

However, the ʈʂa-final stems underwent reanalysis towards r-alternation, which is opposite of what is predicted by lexical statistics. In other words, a purely inductive 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 propose a model of reanalysis that incorporates a markedness bias, and show that it better captures the Malagasy data than an unbiased model.

Note that for the ka-final weak stems, there is also some evidence for markedness influencing reanalysis. Specifically, reanalysis appears to have been partially influenced by h-dissimilation, despite there being no distributional evidence for dissimilation in PMP. However, the lack of evidence makes this pattern harder to confirm. As such, the rest of the paper will focus on markedness effects in ʈʂa-final weak stems, where the effects of markedness are very pronounced.

4 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. In particular, a constraint-based model of reanalysis which incorporates a markedness bias is compared to baseline control models.

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. The model will also make strong, empirically testable predictions about how markedness can influence reanalysis, which can then be applied to other case studies.

The model has three main components. First, it uses Maximum Entropy Harmonic Grammar (MaxEnt; Goldwater & Johnson, 2003; Smolensky, 1986), a probabilistic variant of Optimality Theory. Additionally, to mirror the effect of reanalyses over time, the model will have an iterative (generational) component, in which the output of one iteration of the model becomes the input for the next. 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.

The rest of this section is organized as follows. Section 4.1 outlines the different components of the grammar, including the inputs and constraint set (Section 4.1.1-4.1.3), a procedure for implementing markedness bias (Section 4.1.4-4.1.5), and the iterative component of the model (Section 4.1.6). Finally, Section 4.2 compares the markedness-biased model against several control models, to show that a markedness bias significantly improves model performance.

4.1 Components of a MaxEnt model of reanalysis

Because rates of Malagasy weak stem alternation are probabilistic (as opposed to categorical), I adopt MaxEnt, which uses weighted (instead of ranked) constraints and generates a probability distribution over the set of candidate outputs. In principle, other stochastic inductive models of morphophonological learning, such as the MGL (Section 2.1), would work equally well in matching the Malagasy input data. MaxEnt is adopted because there is existing work on incorporating learning biases in MaxEnt (Wilson, 2006; White, 2013).

Note that unlike classic 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 classic OT, where the winning candidate gets near-perfect probability.

In all subsequent models, constraint weights were learned using the R package *maxent.ot* (Mayer & Zuraw, 2022). Constraint optimization is done using the *optim* function from the R-core statistics library. Constraint weights are restricted to finite, non-negative values.[9]

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, the output is a set of candidates, each with a predicted probability.

4.1.1 Inputs

The input to the model is a set of 1270 nonce weak stem, designed to represent historical Malagasy, presumably before extensive reanalysis had occurred. Relative frequencies of ka, ʈʂ̺a, 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 of number PMP forms available is too few.

For simplicity, only candidates with observed alternants are included in the model. A potential alternate like [p], which is in the Malagasy inventory, but not observed as a weak stem alternant, is assumed to be ruled out by highly weighted faithfulness constraints. In addition, ʈʂ̺a ~ f alternating forms and irregular alternants (e.g. na∼f alternating forms) are excluded, because they are extremely low-frequency and do not influence model outcomes. The input data is summarized in Table 10.

The input matches the *surface* stem allomorphs, and the output candidates are suffixed allomorphs. This is because all reanalyses in Malagasy weak stems are from the non-suffixed to suffixed allomorphs. Reanalysis happens in this direction if speakers have access to the surface stem (or another non-suffixed allomorph), but not the suffixed allomorph. The inputs therefore match the conditions under which speakers would reanalyze weak stems.

This choice of inputs relies on the assumption that the base of reanalysis is *always* a non-suffixed allomorph. A similar approach is taken by Albright (2008; 2010, etc.), who argues that the base of reanalysis is fixed, and is always a single slot of a morphological paradigm. Notably, 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. Fully understanding why the non-suffixed allomorph serves as the base of reanalysis in Malagasy is beyond the scope of this paper. However, the Malagasy data may lead us to rethink Albright's hypothesis that informativeness always determines the base of reanalysis. For Malagasy, one possible alternative factor is the tendency for bases to be isolation stems or other shorter, 'unmarked' forms (Vennemann, 1972; Kuryłowicz, 1945).

[9] Nearly identical results were found using the Excel Solver (Fylstra et al., 1998), which uses the Conjugate Gradient Descent method.

Input	Candidate	Freq	P
ˈvukiʈ͜ʂa	vuˈkiʈ͜ʂana	0	0
	vuˈkirana	56	0.30
	vuˈkitana	131	0.70
ˈvuritra	vuriʈ͜ʂana	0	0
	vuˈrirana	0	0
	vuˈritana	56	1
ˈvukika	vuˈkikana	0	0
	vuˈkihana	490	0.90
	vuˈkifana	57	0.10
ˈvukina	vuˈkinana	440	0.92
	vuˈkimana	40	0.08

Table 10: Sample inputs to the Malagasy model of reanalysis

4.1.2 Faithfulness constraints

The model uses the *Map family of faithfulness constraints, instead of classical feature-based faithfulness constraints (McCarthy & Prince, 1995). *Map constraints, proposed by Zuraw (2010, 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 corresponding 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.[10]

The tableau in (9) demonstrates how *Map violations are assessed for the candidate [ˈvuliʈ͜ʂa]. Candidate (a), where [ʈ͜ʂ] alternates with [t], incurs a violation of *Map(ʈ͜ʂ, t). Meanwhile, candidate (b), where the alternant is [r], incurs a violation of *Map(ʈ͜ʂ,r).

(9)

ˈvuliʈ͜ʂa	*Map(ʈ͜ʂ,t)	*Map(ʈ͜ʂ,r)
a. vuˈlit-ana	*	
b. vuˈlir-ana		*
c. vuˈliʈ͜ʂ-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 [2001]), is a language-specific perceptual map which encodes the perceptual distance between 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 actually 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 different types of learning bias to be incorporated, and have been more successful at modeling phonetic bias in prior work (Wilson, 2006; Hayes & White, 2015).

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

4.1.3 Markedness constraints

The inductive model has four markedness constraints. All four constraints are included because they can be learned simply from local distributional information, and would be learned in comparable inductive models of morphophonological learning.

First, the three markedness constraints *ʈʂ]V, *k]V, and *n]V assess violations for every C]V, where C is at a morpheme boundary. These constraints motivate alternation of the final consonant in weak stems. Reference to morpheme boundaries is necessary because within stems, prevocalic ʈʂ, k, and n are allowed.[11] This approach is similar to the one taken by Pater (2007) and Chong (2020) to explain morphologically-derived environment effects (MDEEs), where static phonotactic patterns mismatch the alternations allowed at morphological boundaries.

The effect of *ʈʂ]V is demonstrated in tableau (10); *k]V and *n]V work in parallel ways. ʈʂa-final weak stems always alternate in the suffixed form. This can be achieved by ranking *ʈʂ]V above competing faithfulness constraints (or by giving *ʈʂ]V a much higher weight). As a result, the faithful candidate (c) is eliminated.

(10)

ˈvuliʈʂa	*ʈʂ]V	*MAP(ʈʂ,t)	*MAP(ʈʂ,r)
a. vuˈlit-ana		∗	
b. vuˈlir-ana			∗
c. vuˈliʈʂ-ana	∗!		

A fourth constraint, *r...r], is used to enforce dissimilation of [r] at the right edge of morpheme boundaries. Again, reference to morpheme boundaries is necessary because within stems, r...r sequences are permitted (e.g. [ˈraraka] 'spilled', [buˈrera] 'weak, limp', [ˈrirana] 'edge'). The effect of *r...r] is demonstrated in tableau (11), where the input stem has a preceding [r], *r...r]. In this tableau, highly ranked *r...r] rules out the r-alternating candidate (b).

(11)

ˈvuriʈʂa	*r...r]	*MAP(ʈʂ,t)	*MAP(ʈʂ,r)
☞ a. vuˈrit-ana		∗	
b. vuˈrir-ana	∗!		∗

The model laid out so far is inductive, and 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.

4.1.4 Learning additional markedness constraints

The central argument of the current study is that reanalysis in Malagasy is partially driven by markedness effects that *cannot* be learned inductively. In this section and the subsequent section, I outline a process for incorporating this markedness component to the model.

First, I propose that the range of markedness constraints that can affect reanalysis are restricted, in that they must already be present in the lexicon. Specifically, the constraint should already be present as a phonotactic tendency in the lexicon. This restriction is consistent with

[11] Examples: be**ʈʂ**oka 'to swell up', **ʈʂ**ano 'box', fo**k**a 'smoke, suck in', a**k**a 'familiar with', a**n**ika 'to climb'

the view that phonotactics guide alternation learning (Tesar & Prince, 2003; Hayes, 2004; Jarosz, 2006), which is supported by experimental evidence (see for example: Pater & Tessier, 2005; Chong, 2021). This restriction also makes empirically testable, language-specific predictions that should be tested in follow-up work, about which markedness effects can affect reanalysis.

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, in order to test for 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. 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 learned the constraint *[+syll][-cont,-voice][+syll], which penalizes intervocalic voiceless stops, and favors [r] over [t] as the alternant for ʈʂa-final weak stems. This contraint therefore motivates reanalysis of t→r. Crucially, it also does not affect the relative preference for different alternants in ka- or na-final weak stems.

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 dissmilation 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.

4.1.5 Incorporating a soft markedness bias

The constraint *V[-cont,-voice]V is added to the model, and assigned a bias towards higher weight. Following Wilson (2006) and White (2017), a bias term, or 'prior', is implemented as a Gaussian distribution over each constraint weight. The bias term, calculated as in (12), is defined in terms of a mean (μ) and standard deviation (σ). For each constraint, *w* is its learned weight, and μ can be thought of as the 'preferred' weight. As such, the numerator of the bias term reflects how much the actual weight deviates from the preferred weight of each constraint, and the penalty resulting from the bias term increases as constraint weights diverge from μ.

(12) $\sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma^2}$

The value of σ^2 determines how much effect the preferred weight (μ) has; lower values of σ^2 result in a smaller denominator, and therefore greater penalty for weights that deviate from their μ. In unbiased models, the goal of learning is to maximize log probability. With the inclusion of the prior, the goal becomes to maximize a different OBJECTIVE FUNCTION, which is the prior term subtracted from the log probability of the observed data.

In principle, both μ and σ^2 can be varied to give constraints a preference towards a certain weight. In the current models, σ^2 is set to fixed values. The markedness constraints *ʈʂ]V, *k]V, *n]V, and *r...r] have no phonotactic support from the lexicon, but are supported by local distributional information. For these constraints, I assume that the weight is learned from the input data, and that the effect of a bias is negligible. This is done by setting σ^2 to an arbitrarily high value (1000).

For the rest of the constraints, σ^2 is set to 0.5, 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 Section 4.2, I provide the specific μ values used for the markedness-biased model, as well as the μ values of baseline control models.

4.1.6 Using an iterative model

To simulate reanalysis over time, I use a generational model, in which the output of one iteration of the model becomes the input to the next iteration. Similar models of language change are by no means new, and there are various approaches to doing so. For example, 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), or as the result of competing grammars (Yang, 1976). More recently, Zuraw (2000), like the current study, uses a probabilistic variant of OT to model change over generations of speakers.

The model adopted in the current paper is simpler than many existing generational learning models. It does not account for individual variation, or factors such as usage frequency, which is known to affect the likelihood of a form to be reanalyzed (e.g. Mańczak, 1980; Bybee, 2003). Nevertheless, the current implementation suffices for capturing how reanalysis is in the direction of high-probability alternants.

The model works as follows: MaxEnt generates a probability distribution over the set of candidate outputs. For each input, the winning candidate is selected by sampling from the candidate set, using this probability distribution. The winning candidate is then used as the input to the next iteration of model fitting.

Under this approach, candidates with low-probability alternants are more likely to undergo reanalysis. For example, a hypothetical stem-suffix pair [ˈvulika]∼[vuˈlif-ana], where final [ka] alternates with the low-probability alternant [f], is likely to be reanalyzed as [ˈvulika]∼[vuˈlih-ana]. In contrast, a stem which already undergoes ka∼h alternation is less likely to be reanalyzed.

Because random sampling causes each iteration of the model to vary slightly, all subsequent models were run 10 times, and predicted probability values are the mean of these 10 trials.

4.2 Model comparison

This section compares markedness biased models against controls, to evaluate the effect of markedness in improving model predictions. Although it is not the focus of the current paper, models with a p-map bias are also tested. These models are included to confirm that markedness effects improve model predictions after controlling for perceptual similarity effects, which have been substantiated by prior research (White, 2013, 2017).

A total of four models are compared: the first two conditions, FLAT-PRIOR and P-MAP, are controls. They are compared to two conditions with a markedness bias, labeled MARKEDNESS and FULL (which includes both a markedness and p-map bias). The priors assigned to each condition are explained below, and summarized in Table 11.

If reanalysis is in fact driven by a markedness bias in Malagasy, then the MARKEDNESS and FULL models should outperform their respective control conditions, FLAT-PRIOR and P-MAP. If, instead, reanalysis is rooted in a p-map bias, adding a markedness bias should not improve model fit. Instead, the P-MAP condition (and potentially the FULL condition) should perform better than the FLAT-PRIOR model, and the FULL condition should not perform better than the P-MAP condition.

P-MAP condition (control). The p-map condition (labeled P in Table 11) 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.[12] As there are no confusability experiments for Malagasy, I use results from Warner et al. (2014), a study of consonant confusability in English, as a proxy.[13] 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].

FLAT-PRIOR condition (control). The FLAT-PRIOR model (labeled FLAT in Table 11) is also a control condition. In this condition, every constraint with a bias term has the same μ of 3.3, which is the mean of all μ values assigned to the *MAP constraints in the P-map condition. This condition will be compared against the MARKEDNESS condition. It is included because as discussed in White (2013), a model with uniform (but non-zero) μ values is a better control than a model with no bias terms at all.

MARKEDNESS condition. The MARKEDNESS condition (labeled M in Table 11) assigns a uniform prior, $\mu = 3.3$, to all faithfulness constraints. The markedness constraint *V[-cont,-voice]V is assigned a high prior ($\mu = 7$). This value is higher than the μ assigned to the competing faithfulness constrain *MAP(ʈʂ,r), but is otherwise arbitrary. This condition differs from the FLAT-PRIOR condition *only* in the μ value assigned to *V[-cont,-voice]V; the two models are otherwise identical.

FULL condition. The FULL condition has both a markedness bias and a p-map bias. Like the MARKEDNESS condition, *V[-cont,-voice]V is assigned a μ value of 7. The P-MAP and FULL conditions are identical except for the μ values assigned to *V[-cont,-voice]V.

4.2.1 Model results after one iteration

Table 12 shows results after one model iteration. The column titled 'Obs (PMP)' shows the observed probability of the input candidates, and reflects the distribution of alternants before reanalysis. The column 'Obs (Mal)' reflects the distribution of alternants in modern Malagasy, *after* reanalysis. Due to reanalysis of ʈʂa-final forms in the direction of t→r (see §3), modern Malagasy shows a much higher rate of ʈʂa∼r alternation than PMP.

[12] 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).

[13] I use Warner et al. (2014) because unlike other studies of English consonant confusability (e.g. Wang & Bilger, 1973; Cutler et al., 2004), it tests for confusability of phonemes with [ɾ].

		μ			
Constraint	σ^2	FLAT	M	P	FULL
*ʈ͜ʂ]V	1000	0	0	0	0
*k]V	1000	0	0	0	0
*n]V	1000	0	0	0	0
*r...r]	1000	0	0	0	0
*MAP(tr,r)	0.5	3.3	3.3	5.13	5.13
*MAP(tr,t)	0.5	3.3	3.3	2.82	2.82
*MAP(n,m)	0.5	3.3	3.3	1.83	1.83
*MAP(k,f)	0.5	3.3	3.3	2.76	2.76
*MAP(k,h)	0.5	3.3	3.3	0	0
*V[-cont,-vc]V	0.5	3.3	7	0	7

Table 11: Constraints and bias terms by condition (P = p-map condition, M = markedness condition)

Results in the control conditions (FLAT-PRIOR and P-MAP) are comparable. Both match the frequencies of the input data closely. The two conditions with a markedness bias perform essentially the same. Both predict an increase in the probability of [vuˈkirana] (by 6%), and therefore reanalysis to be in the direction of t→r. In other words, adding a markedness bias does appear to improve model predictions. The magnitude of change is relatively small after one iteration of the model. However, as seen in the following section, the model will approach the distribution seen in modern Malagasy after multiple iterations.

				Predicted			
Input	Cand	Obs (PMP)	Obs (Mal)	FLAT-PRIOR	P-MAP	MARK	FULL
vukiʈ͜ʂa	vukirana	0.30	0.95	0.31	0.31	**0.36**	**0.36**
	vukitana	0.70	0.05	0.69	0.69	**0.64**	**0.64**
	vukiʈ͜ʂana	0	0	0	0	0	0
vuriʈ͜ʂa	vurirana	0	0	0	0	0.04	0
	vuritana	1	1	1	1	0.96	1
	vuriʈ͜ʂana	0	0	0	0	0	0
vukika	vukikana	0	0	0	0	0	0
	vukihana	0.90	0.95	0.90	0.90	0.90	0.90
	vukifana	0.10	0.05	0.10	0.10	0.10	0.10
vukina	vukinana	0.92	0.98	0.91	0.91	0.92	0.93
	vukimana	0.08	0.02	0.09	0.09	0.08	0.08

Table 12: Predicted probability of models after one iteration (mean of 10 trials)

4.2.2 Model results after 10 iterations

After 10 iterations, the models with a markedness bias clearly outperform corresponding control conditions. This is seen in Table 13, which shows the proportion of variance explained (R^2) and log likelihood ($\hat{L}$) for each model after 10 iterations, where the model predictions are fit to the modern Malagasy distribution. Additionally, Fig. 3 compares the model fit (R^2) in the four conditions over 10 iterations.

Condition	$\mathbf{R^2}$	**($\hat{L}$)**
FLAT	0.60	-9273
P-MAP	0.58	-9618
MARKEDNESS	0.99	-6460
FULL	0.98	-6233

Table 13: Results after 10 iterations: Proportion of variance explained (R^2) and log likelihood ($\hat{L}$), of model predictions fit to modern Malagasy

Fig. 3 shows that for the two control models (FLAT-PRIOR and P-MAP), model fit does not improve over iterations ($R^2 \approx 0.6$). In contrast, both the MARKEDNESS and FULL are able to account for over 98% of the variation in the observed Malagasy data by the 10th iteration($R^2 \geq 0.98$). As seen in Table 13, MARKEDNESS and FULL also perform better than the controls in terms of log-likelihood.

Adding a p-map bias does not strongly affect model fit, as the FLAT-PRIOR and P-MAP conditions perform similarly poorly. However, for log-likelihood, the FULL model ($\hat{L} = -6233$) performs slightly better than the MARKEDNESS model ($\hat{L} = -6460$). In other words, adding a p-map bias on top of a markedness bias does slightly improve model fit.

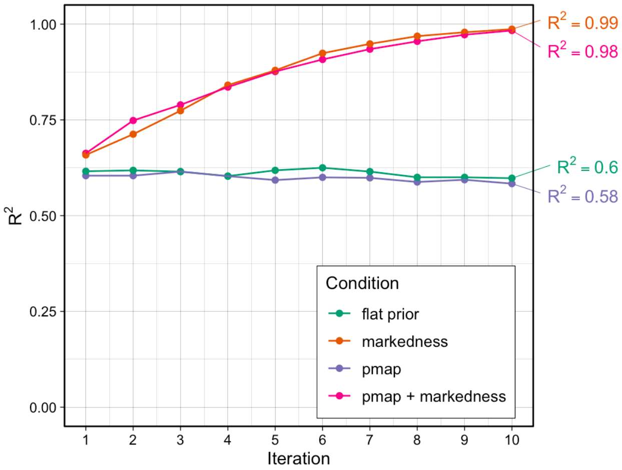

Figure 3: Model fit by conditions over 10 iterations (mean of 10 trials)

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. 4, which plots the change in predicted probabilities over 10 trials, 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 non-alternating candidate ([ˈvukiʈʂa∼vuˈkiʈʂana]) is not shown, but is never observed, and is consistently assigned zero probability by all models.

In the two conditions with a markedness bias, the model successfully predicts an increase in the ʈʂ∼r alternating candidate [vuˈkiʈʂa∼vuˈkir-ana], and therefore closely matches the Malagasy data. At the same time, for the input ˈvuriʈʂa, where r-dissimilation should block the r-alternating candidate, the models consistently predicts the correct candidate [vuˈritana] (P = 1).

For ka- and na-final weak stems, all four models perform similarly well. This is demonstrated in Fig. 5, which plots the change in predicted probabilities over 10 trials, for ka- and na-final weak stems. All four conditions assign high probabilities to the h-alternating candidate for ka-final weak

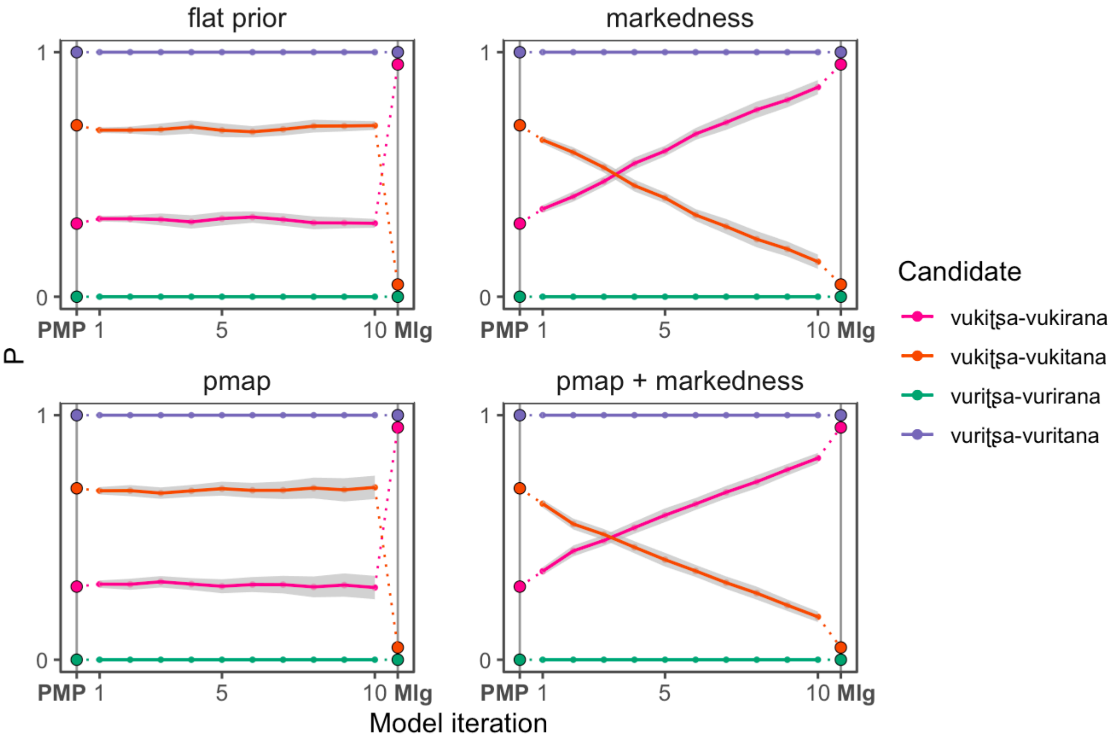

Figure 4: Predicted probabilities of candidates over 10 iterations for ʈʂa-final weak stems (mean of 10 trials). Grey intervals indicate standard error, and observed rates of alternation in PMP and Malagasy are given for reference.

stems, and the non-alternating candidate for na-final weak stems. The MARKEDNESS and FULL conditions do assign a slightly lower probability to the preferred candidates][(ˈvukika∼vuˈkihana], [ˈvukina∼vuˈkinana]), but the magnitude of this change is very small.

Table 14 shows the detailed predictions of each condition on the 10th iteration. The two control models (FLAT-PRIOR and P-MAP) largely match the input (historical PMP) distribution, and therefore under-predict rates of ʈʂa∼r alternation. Both the MARKEDNESS and FULL conditions predict a large magnitude of reanalysis in the direction of t→r, and assign the r-alternating candidate (vuki**r**ana) a high probability ($P_{\text{MARKEDNESS}} = 0.86$; $P_{\text{FULL}} = 0.82$). The MARKEDNESS model does slightly better, and predicts a higher rate of ʈʂa-r alternation.

Intriguingly, for ka-final stems, the FULL model actually does slightly better than the MARKEDNSS model at matching modern Malagasy. Specifically, the MARKEDNESS model predicts a lower rate of the preferred ka∼h alternating candidate. This is because in the FULL condition, the p-map bias motivates higher rates of ka∼h alternation (as *MAP(k,h) has a lower μ than *MAP(k,f)). In contrast, the MARKEDNESS condition assigns all faithfulness constraints the same prior. As a result, there is pressure for candidates to have the same rate of alternation, all else held equal.

Overall, model results support the hypothesis that reanalysis in Malagasy weak stems is largely driven by a markedness bias which penalizes intervocalic stops. Additionally, comparison of the MARKEDNESS and FULL models shows that although a perceptual bias does not noticeably improve model fit, it does improve the model predictions for ka-final weak stems.

In summary, adding a markedness bias improves model fit in both the MARKEDNESS and FULL conditions. Both models predict renalaysis of t→r for ʈʂa-final weak stems, while matching the frequencies of the input data for ka- and na-final weak stems.

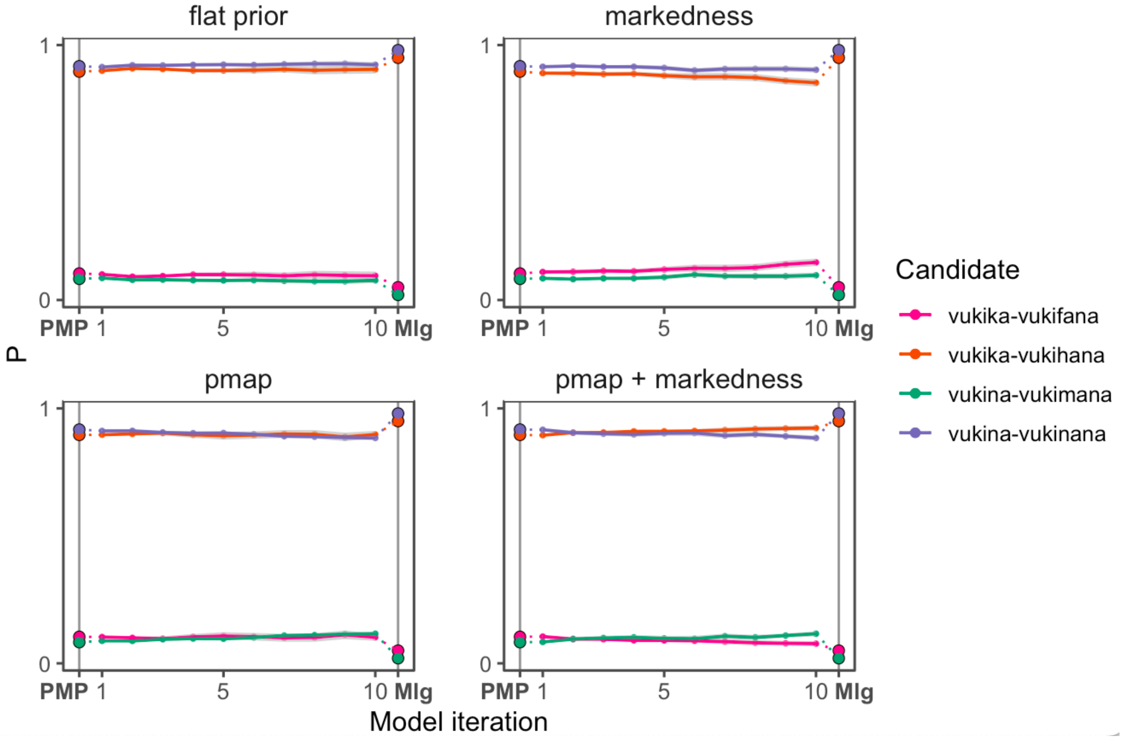

Figure 5: Predicted probabilities of candidates over 10 iterations for ka- and na- weak stems.

				Predicted			
Input	Cand	Obs (PMP)	Obs (Mal)	FLAT-PRIOR	PMAP	MARK	FULL
vukitra	vukirana	0.30	0.95	0.30	0.30	0.86	0.82
	vukitana	0.70	0.05	0.70	0.70	0.14	0.18
	vukitrana	0	0	0	0	0	0
vuritra	vurirana	0	0	0	0	0.24	0
	vuritana	1	1	1	1	0.76	1
	vuritrana	0	0	0	0	0	0
vukika	vukikana	0	0	0	0	0	0
	vukihana	0.90	0.95	0.90	0.90	0.85	0.92
	vukifana	0.10	0.05	0.10	0.10	0.15	0.08
vukina	vukinana	0.92	0.98	0.92	0.88	0.90	0.90
	vukimana	0.08	0.02	0.08	0.12	0.10	0.10

Table 14: Predicted probability of models after 10 iterations (mean of 10 trials)

4.3 Generational models and the choice of σ^2

In the current model, σ^2 is set to 0.5, which allows for the bias to have a small magnitude of effect that adds up over multiple iterations. By the 10th iteration, the model closely matches the rates of alternation observed in modern Malagasy.

A superficially similar outcome can be achieved by removing the generational component of the model, and simply setting σ^2 to a lower value. A lower σ^2 allows the bias to have a stronger effect, so that the model predicts a greater magnitude of change in just one iteration. Fig. 6

shows the model fit over 10 iterations for the FULL model when σ^2 is varied, and μ values are held constant. Both the high-sigma model ($\sigma^2 = 0.5$) and low-sigma model ($\sigma^2 = 0.1$) converge on the same outcome, but the low-sigma model does so much faster, after just 1-2 iterations.

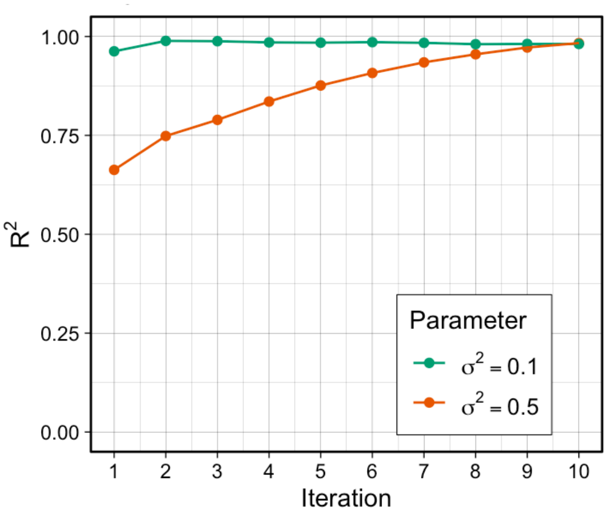

Figure 6: R^2 over 10 iterations of the FULL model, when σ^2 is varied

Although a low-sigma model achieves the same outcome as a multigenerational high-sigma model, I argue that the multi-generational model is preferable for the following reasons. First, it is conceptually more plausible that reanalysis happens gradually. This is especially true for a case like Malagasy, where the reanalysis of t→r cannot be attributed to sound change, and both alternants are phonemic.

A generational model also predicts randomness and variation; in the current paper, this comes from randomly sampling the winning candidate that becomes the input to the next model iteration. This 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.

5 Conclusion

The current paper looked at reanalysis in Malagasy weak stems, and found that for the ʈʂa-final stems, the direction of reanalysis cannot be predicted by local distributional information. Instead, I argue that reanalysis of t→r is motivated by a markedness constraint against intervocalic (voiceless) stops. This markedness constraint is typologically well-motivated, and also present in the Malagasy lexicon as a phonotactic tendency. Based on these results, I outline a model of reanalysis with a markedness learning bias. This model outperformed control models, and was able to closely match the Malagasy data.

The approach to incorporating markedness laid out in this study makes empirical predictions about which markedness effects can affect reanalysis. Specifically, I argue that the markedness effects affecting reanalysis are restricted, and must already present in a language's phonological grammar. In the case of Malagasy, the relevant constraint *V[-cont,-voice]V was found to have significant weight in a phonotactic grammar.

Finally, a model which fully captures reanalysis would be more complex than the one developed here, and should be explored in future work. For one, the current model ignores factors such as usage frequency (Bybee, 2003), and assumes that bias factors remain the same over iterations of the model. In addition, the current model assumes surface-base representations, where surface stem allomorphs are the inputs. However, reanalysis in Malagasy is also potentially compatible with a model of base competition, in which outputs are faithful to multiple listed allomorphs, but also sensitive to markedness effects (Breiss, 2021). Future work will consider how different parameters can be varied in modeling reanalysis, as well as how input forms should be represented.

References

Adelaar, Alexander (1989). Malay influence on Malagasy: Linguistic and culture-historical implications. *Oceanic Linguistics, A Special Issue on Western Austronesian Languages (Summer, 1989)* **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 (2008). A restricted model of UR discovery: Evidence from Lakhota. *Ms, MIT* .

Albright, Adam (2010). Base-driven leveling in Yiddish verb paradigms. *NLLT* **28:3**. 475–537.

Albright, Adam & Bruce Hayes (2002). Modeling English past tense intuitions with minimal generalization. In *Proceedings of the ACL-02 workshop on morphological and phonological learning*, 58–69.

Albright, Adam & Bruce Hayes (2003). Rules vs. analogy in past tenses: A computational/experimental study. *Cognition* **90:2**. 119–161.

Albright, Adam C (2002). *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.

Blust, Robert & Stephen Trussel (2010). Austronesian comparative dictionary, web edition. *Blust's Austronesian Comparative Dictionary Website* .

Breiss, Canaan (2021). *Lexical conservatism in phonology: theory, experiments, and computational modeling*. PhD dissertation, University of California, Los Angeles.

Bybee, Joan (2003). *Phonology and language use (cambridge studies in linguistics)*. Cambridge: Cambridge University Press.

Cheung, Hintat (2000). Three to four-years old children's perception and production of Mandarin consonants. *Language and Linguistics* **1:2**. 19–38.

Chong, Adam J (2020). Exceptionality and derived-environment effects: A comparison of Korean and Turkish. *Phonology* .

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

Colantoni, Laura & Jeffrey Steele (2005). Liquid asymmetries in French and Spanish. *Toronto Working Papers in Linguistics* **24**.

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.

Dahl, Otto Christian (1951). Malgache et Maanjan: une comparaison linguistique. *Avhandlinger utgitt av Egede Instituttet* .

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.

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.

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*, vol. 111120, .

Hayes, Bruce (2004). Phonological acquisition in optimality theory: the early stages. In *Constraints in phonological acquisition*, 158–203. Cambridge University Press.

Hayes, Bruce (2011). *Introductory phonology*, vol. 32. John Wiley & Sons.

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. *LI* **39:3**. 379–440.

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* .

Jarosz, Gaja (2006). *Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory*. PhD dissertation, ohns Hopkins University.

Jun, Sun-Ah (1994). The status of the lenis stop voicing rule in Korean. In YoungKey Kim-Renaud (ed.), *Theoretical issues in Korean linguistics*, 101–114. Stanford:CSLI.

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.

Keenan, Edward L & Maria Polinsky (2017). Malagasy (austronesian). *The handbook of morphology* 563–623.

Kiparsky, Paul (1965). *Phonological change*. PhD dissertation, MIT.

Kiparsky, Paul (1968). Linguistic universals and linguistic change. In Emmon W. Bach & Robert T. Harms (eds.), *Universals in linguistic theory*, 170–202. New York: Hold, Rinehart & Winston.

Kiparsky, Paul (1978). Analogical change as a problem for linguistic theory. *Studies in the Linguistic Sciences Urbana, Ill* **8:2**. 77–96.

Kirchner, Robert M. (1998). *An effort-based approach to consonant lenition*. PhD dissertation, University of California, Los Angeles.

Kuryłowicz, Jerzy (1945). La nature des procès dits «analogiques». *Acta linguistica* **5:1**. 15–37.

de La Beaujardière, Jean-Marie (2004). Malagasy dictionary and encyclopedia of Madagascar.

Lavoie, Lisa M (2001). *Consonant strength: Phonological patterns and phonetic manifestations*. Routledge.

Lewis, Paul M., Gary F. Simons & Charles D. Fennig (2014). *Ethnologue: Languages of Asia*. SIL international.

Mahdi, Waruno (1988). *Morphophonologische Besonderheiten und historische phonologie des Malagasy*, vol. 20. D. Reimer.

Mańczak, Witold (1980). Laws of analogy. In Jacek Fisiak (ed.), *Historical morphology*, 183–188. Berlin: de Gruyter.

Mayer, Connor & Kie Zuraw (2022). *maxent.ot: Perform maxent optimality theory analyses in r*. http://connormayer.com/maxent_ot.html. R package version 0.0.0.9000.

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.

Moreton, Elliott & Joe Pater (2012). Structure and substance in artificial-phonology learning, part I: Structure. *Language and linguistics compass* **6:11**. 686–701.

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

Niyogi, Partha (2006). *The computational nature of language learning and evolution*. MIT press Cambridge, MA.

Nosofsky, Robert M (2011). The generalized context model: An exemplar model of classification. *Formal approaches in categorization* 18–39.

O'Neill, Timothy (2015). *The phonology of Betsimisaraka Malagasy*. PhD dissertation, University of Delaware.

Pater, Joe (2007). The locus of exceptionality: Morpheme-specific phonology as constraint indexation. In *University of Massachusetts occasional papers 32: Papers in optimality theory iii*, 259–296. Amherst: GLSA.

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.

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

Rasoloson, Janie & Carl Rubino (2005). Malagasy. In Alexander Adelaar & Nikolaus Himmelmann (eds.), *The austronesian languages of asia and madagascar*, 456–488. Routledge Abingdon.

Sapir, Edward (1915). Notes on Judeo-German phonology. *The Jewish quarterly review* **6:2**. 231–266.

Selkirk, Elisabeth (1980). Prosodic domains in phonology: Sanskrit revisited. In Mark Arnoff & Mary-Louise Kean (eds.), *Juncture*, 107–129. Anma Libri.

Singleton, Jenny L & Elissa L Newport (2004). When learners surpass their models: The acquisition of American Sign Language from inconsistent input. *Cognitive psychology* **49:4**. 370–407.

Smolensky, Paul (1986). Information processing in dynamical systems: Foundations of harmony theory. Tech. rep. Colorado Univ at Boulder Dept of Computer Science.

Steriade, Donca (2009 [2001]). 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.

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.

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.

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* The phonology of the world's languages. Oxford University Press.

White, James (2013). *Bias in phonological learning: Evidence from saltation.* PhD dissertation, University of California, Los Angeles.

White, James (2017). Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Lg* **93:1**. 1–36.

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

Yang, Hsiu-fang (1976). The phonological structure of the Paran dialect of Sediq. *Bulletin of the Institute of History and Philology Academia Sinica* **47:4**. 611–706.

Zuraw, Kie (2000). *Patterned exceptions in phonology*. PhD dissertation, University of California, Los Angeles.

Zuraw, Kie (2010). 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.