Markedness bias in the reanalysis of Samoan thematic consonant alternations

Jennifer Kuo

[a] *Cornell University, Dept of Linguistics, Ithaca, NY, US*

Abstract

Paradigms with conflicting data patterns can be difficult to learn, resulting in a type of language change called *reanalysis*. Existing models of morphophonology predict reanalysis to occur in a way that matches frequency distributions within the paradigm. Using evidence from Samoan, this paper argues that instead, reanalysis is sensitive to both frequency and the reduction of markedness. More concretely, I find that reanalysis of Samoan thematic consonants is generally towards the historically more frequent alternants (in line with a frequency-matching approach), but is also modulated by OCP-place effects. These results are confirmed in an iterated learning model that is based in MaxEnt (Goldwater and Johnson, 2003). Additionally, I propose that markedness effects must be i) already active in stem phonotactics (*active markedness* restriction), and ii) phonetically motivated (*phonetic naturalness* restriction). The Samoan data is compatible with these restrictions; OCP-place is active in Samoan stem phonotactics, supporting the active markedness restriction. Additionally, in a study where phonetic similarity is measured as the spectral distance between two phones, I find that similarity of consonants is closely correlated with the strength of OCP-place effects in Samoan; this suggests that OCP-place is rooted in phonetic similarity avoidance, supporting the phonetic naturalness restriction.

Keywords: morphophonology, Samoan, markedness, bias learning, OCP-place, phonetic similarity

1. Introduction

Particularly since Kiparsky (1965, 1997, 1978, et seq.), it has been recognized that *language change* can serve as a robust "natural laboratory" for understanding how children learn and mislearn patterns outside the constraints of a laboratory setting. In this paper, I take this approach and focus on a specific type of language change I refer to as *reanalysis* to gain insight into the learning of morphophonological paradigms.

Paradigms can often have conflicting data patterns. Consider the case of English past tense formation, where past tense can be formed in multiple ways (e.g. *want/wanted*, *bleed/bled*, *speak/spoke*, etc.). This is potentially challenging for learners, who, when presented with a novel word, are faced with conflicting data patterns about how to form the past tense. For example, given a hypothetical stem like *gleed*, the learner has multiple choices for the past tense, a subset of which are given in Table 1.

This type of ambiguity can be challenging for learners, resulting in acquisition errors (e.g. *go/goed* instead of *go/went*) in English. Such errors can be adopted into the speech community, resulting in a type of change over time I refer to as *reanalysis*. Some exam-

Output	**Generalization**	**Examples**
gleeded	∅→ əd / {d, t} ___	*want, need, start, decide*
gled	i→ ɛ / {l,r}___d	*read, lead, bleed, breed*
glode	i→ o / C___[+cons]	*speak, freeze, weave*
gleed	no change / ___{d,t}	*shed,spread,put*

Table 1: English past tense formation for *gleed* (Albright and Hayes, 2003, p. 128)

ples of reanalysis in English past tense include *help/halp→help/helped* (∼1300, OED) and *dive/dived→dive/dove* (∼1800, OED)[1]

Existing models of morphophonology, such as the Minimal Generalization Learner (Albright and Hayes, 2003), predict reanalysis to be frequency-matching, occurring in a way that matches probabilistic distributions of the input data. In this paper, I propose that in fact, reanalysis responds to two factors: both frequency and the reduction of markedness.

In this paper, I investigate this issue using evidence from reanalysis in a set of Samoan consonant alternations. I present data suggesting that while reanalysis is generally in the direction predicted by frequency-matching, it is also modulated by effects of OCP-place, or the Obligatory Contour Principle on place of articulation (Greenberg, 1950; McCarthy, 1988). These results are confirmed using iterated learning models that simulate the cumulative effect of reanalyses over time. In particular, I find that models with a markedness bias outperform purely frequency-matching models, showing evidence for markedness effects in reanalysis.

I further propose that the markedness effects present in reanalysis are constrained by two factors: stem phonotactics and phonetic naturalnesss. First, the markedness effects allowed to affect reanalysis must already be active in the language's root phonotactics. Moreover, when speakers draw on phonotactics to aid in the learning of alternations, they are extracting principles rooted in phonetic naturalness, rather than any arbitrary statistical generalizations active in the phonotactics. As a preview of results, I find that the Samoan data is consistent with both restrictions.

The rest of this paper is organized as follows: Section 2 discusses the literature on models of reanalysis and learning biases. Following this, Section 3 gives the descriptive facts of Samoan thematic consonant alternations, which are the empirical focus of the current study. In Section 4, I present evidence that OCP-place is both active in Samoan phonotactics and phonetically motivated (i.e. can be characterized as phonetic similarity avoidance). Finally, Section 5 presents evidence that reanalysis in Samoan is sensitive to OCP-place effects, and 6 confirms these results with a quantitative model of reanalysis.

2. Background

2.1. Frequency-matching in models of morphophonological learning

When speakers are faced with variable patterns in a morphophonological paradigm, they are known to **frequency-match**, applying these patterns in a way that matches the proportion at which they occur in the data. For example, Ernestus and Baayen (2003) study Dutch final devoicing, where final obstruents are devoiced, results in voicing alternations such as in (1).

(1) *Dutch voicing alternations*

[vɛrˈʊɛi**t**]	[vɛrˈʊɛi**d**en]	'widen'
[vɛrˈʊɛi**t**]	[vɛrˈʊɛi**t**en]	'reproach'

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

Across 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 and Baayen (2003) find that when speakers are told to provide the suffixed form of nonce stems, they apply voicing alternations in a way that aggregately matches these distributional patterns.[2]

Frequency-matching has been found to predict adult linguistic behavior in various other experiments, including: Eddington (1996, 1998, 2004); Coleman and Pierrehumbert (1997); Berkley (2000a); Zuraw (2000); Bailey and Hahn (2001); Frisch and Zawaydeh (2001); Albright (2002b); Albright and Hayes (2003); Hayes and Londe (2006); Hayes et al. (2009); Pierrehumbert (2006); Jun and Lee (2007). Sociolinguistic studies also demonstrate that children frequency-match adult speech patterns (Labov, 1994, Ch. 20).

Existing quantitative models of reanalysis (and more generally, models of morphophonological learning) are all frequency-matching. These models include neural networks (Rumelhart and McClelland, 1987; MacWhinney and Leinbach, 1991; Daugherty and Seidenberg, 1994; Hare and 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 and Marinov, 1993).

One representative model of this variety is the Minimal Generalization Learner (MGL; Albright and Hayes, 2003; Albright, 2002b, 2010, etc). The MGL is a rule-based model; it compares members of a paradigm and from here, iteratively learns rules of increasing generality. The result is a system of stochastic rules which predict the inflected form of a paradigm given an input base. 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 successful at predicting the direction of reanalysis in many languages, including Lakhota (Albright, 2002a), Yiddish (Albright, 2010), and Korean (Kang, 2006). However, like other models of morphophonology, it is frequency-matching and directly derives its predictions from statistical distributions within the paradigm.

2.2. Learning biases

In many cases, such as with the Dutch voicing alternations discussed above, frequency-matching 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 2023a) or under-learn them (reflecting a bias against the target pattern; e.g. Hayes et al. 2009; Becker et al. 2011).

Two types of bias have been discussed in the literature: complexity bias, or a bias against formally complex patterns (Moreton and Pater, 2012a), and substantive bias, or a bias against phonetically unnatural patterns (Moreton and 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). Another possibility, which is the empirical focus of this paper, is a so-called **markedness bias**, or a bias against surface forms that dispreferred on phonetic grounds. Specifically, I argue that reanalysis in Samoan is sensitive to transvocalic OCP-place,

[2]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 and Newport, 2005, 2009; Schumacher and Pierrehumbert, 2021).

which is a surface constraint against sequences of homorganic consonants separated by intervening vowel(s).

2.3. Phonetic naturalness and the OCP

Transvocalic OCP-place, which bans (or penalizes) sequences of homorganic consonants separated by intervening vowels, is 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 and Pater, 2006, 2008), English (Berkley, 1994, 2000b), Tigrinya (Buckley, 1997), Japanese (Kawahara et al., 2006), and Chol (Gallagher and Coon, 2009).

Notably, 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, 1991, 1995; Wilson and Obdeyn, 2009). In Arabic, for example, OCP-place effects are stronger for coronals that share the same sonorancy specification (Pierrehumbert, 1993; Frisch and Zawaydeh, 2001). More concretely, sequences like [t...d] and [n...l] are more marked than [t...l] and [n...d].

Frisch (1996) and Frisch et al. (2004) argue that the gradience of OCP-place provides strong evidence that there is a functional phonetic motivation for the constraint. In particular, speakers tend to avoid sequences of phonetically similar sounds due to general processing constraints that disfavor repetition; evidence for this kind of processing constraint has been found in psycholinguistics, in work such as Dell (1984) and Sevald and Dell (1994).

2.4. Active vs. universal markedness

When we consider markedness effects in reanalysis, it is important to look at whether such effects are subject to language-specific constraints.

One view, which I refer to as 'universal markedness', is that all possible markedness constraints as defined by Universal Grammar can affect reanalysis. These effects may be rooted in phonetic naturalness (as was argued for OCP-place above), but are not otherwise constrained. Another more restrictive view, which I call 'active markedness',' is 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 for several reasons. First, it ties into 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 and 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 underattested.

The active markedness approach also ties into many theories of acquisition, which argue that phonotactics are learned before alternations and aid in the later learning of alternations (Hayes, 2004; Jarosz, 2006; Tesar and Prince, 2003; Yang, 2016). In fact, various experimental work supports the idea that phonotactics aids in alternation learning. For example, Pater and 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. Martin presents evidence that the same phonotactic constraints present within morphemes are also active as a weaker, gradient effect across morpheme boundaries. In other words, there is evidence that speakers generalize phonotactic constraints across morpheme boundaries. Martin's empirical focus was on compound formation, but it is conceivable that stem-internal phonotactics could also constrain alternations.

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. In Section 4.1, I show that OCP-place is active in Samoan stem phonotactics, in line with the active markedness proposal.

3. Samoan thematic consonant alternations

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 across 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 has so-called thematic consonant alternations, where under suffixation, a consonant of unpredictable quality may surface, as seen by the examples in (2) using the ergative suffix. These are the focus of the current paper, and will be described in more detail below.

(2) *Examples of thematic consonant alternations in Samoan*

stem	stem+ERG	gloss
e?e	e?e**t**ia	'be raised'
ala	ala**f**ia	'path, way'
tautau	tautau**l**ia	'to hang up'

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 and 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 and Bradshaw (2013) on stem phonotactics. Moore-Cantwell (2008) was the first 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; Hovdhaugen et al., 1986; Greenhill and Clark, 2011). Historical comparative data is available in both the Austronesian Comparative Dictionary (ACD; Blust et al., 2023) and the Polynesian Lexicon Project (POLLEX; Greenhill and Clark, 2011).

In the rest of this section, I give an overview of Samoan phonology and describe the regular sound changes that lead to the development of thematic consonant alternations in Samoan. Unless otherwise noted, descriptive generalizations are taken from Mosel and Hovdhaugen (1992).

3.1. Phoneme inventory and phonotactics

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, but native speakers are generally cognizant of both levels. The descriptions in this section, as well as 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 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 contrast. 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, as they behave like long vowels in stress assignment. Note that Mosel and Hovdhaugen (1992) propose a larger set of diphthongs that additionally includes /eo, oi, ui/.

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

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

The consonant inventory (of the *tautala lelei* register) is given in (3). /ʔ/ is phonemic, but described by Mosel and 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.

3.2. Samoan thematic consonants and their historical development

In Samoan, thematic consonant alternations are observed in a variety of suffixal contexts, listed in (2). Where thematic consonants surface in the examples, they are shown in boldface. Of these, /-(C)i/ is non-productive and only observed in lexicalized forms (Mosel and Hovdhaugen, 1992, p. 205). For the rest of this chapter, I focus on just the **ergative suffix**, since it is the most productive of the suffixes in (2).

The ergative suffix has the allomorphs /-a/, /-ina/, /-ia/, /-Cia/, and /-na/, where 'C' can be one of the consonants /f, m, t, s, l, ŋ, ʔ/. Examples of each allomorph are given in Table 3. 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. Note that /-ia/, which is also a vowel-initial allomorph, is non-productive and relatively infrequent in Samoan. The consonant-initial allomorphs /-Cia/ and /-na/ 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's (1966) Samoan dictionary, 34% (n=179/527) have a thematic consonant.

Thematic consonants arose as a result of a historical process of final consonant deletion, which affected many Oceanic languages, including all languages in the Polynesian subfamily. For the affected languages, stem-final consonants were maintained in suffixed forms but lost in unsuffixed forms, resulting in ∅∼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*).

Assuming no reanalyses, stems that historically ended in vowels (and in some consonants,

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’
?	=(C)i	sua/sua**t**i ‘dig up (violently)’ sulu/sului “dried banana-leaf”
intensifier	-(C)aʔi	oso/oso**f**aʔi “jump/attack” ufi/ufiaʔi “cover/covered”

Table 2: Samoan suffixes with thematic consonant alternations

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	*pulu**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 3: Samoan thematic consonant alternations

as summarized in Table 4) should take a vowel-initial suffix. Otherwise, the suffix that surfaces is of the form /-Cia/, where /C/ is a reflex of the POc stem-final consonant. 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 (e.g. ua∼ua-i**n**a <*qusa**n** ‘to rain’), and /-na/ surfaces elsewhere. Note that /-ina/ is homophonous with the vowel-initial /-ina/ allomorph.[3]

Finally, Table 4 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. This is demonstrated in the rightmost column of Table 3, which shows the POc reconstruction corresponding to each Samoan stem-suffix pair.

However, it turns out that there is substantial mismatch between the POc final consonant, and the ergative allomorph which surfaces in modern Samoan. These mismatches suggest that extensive reanalyses have occurred. As previewed above, I argue that reanalyses are sensitive to both statistical distributions within paradigms, but also to markedness effects, specifically OCP-place. The following section discusses the basis for OCP-place effects in Samoan reanalysis, while Section 5 goes into more detail about the observed patterns of reanalysis.

[3]Allomorphy of /-a/, /-ia/, and /-ina/ is not the focus of the current paper, but is thought to have resulted from regular sound changes between POc and Proto-Polynesian (Churchward, 1951; Pawley, 2001)

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

Table 4: Samoan reflexes of POc final consonants[1]
[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.

4. OCP-place in Samoan

I propose that markedness effects present in reanalysis must be both phonetically natural and active in the language's stem phonotactics. In this section, I show that Samoan is consistent with both restrictions. First, as will be discussed in Section 4.1, OCP-place is present in Samoan stem phonotactics. Additionally, based on an acoustic study of Samoan (Section 4.2), I find that the strength of OCP-place effects directly correlate with the phonetic similarity of target segments; these results support Frisch's proposal that OCP-place is phonetically grounded and should be characterized as similarity avoidance.

4.1. OCP-place in Samoan stem phonotactics

OCP-place effects have been documented across multiple Polynesian languages (Krupa, 1966, 1967, 1971). Alderete and 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 *[**f**u**m**a]). 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. *[**nul**a] is marked because [n] and [l] are both coronal sonorants).

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

4.1.1. Data and basic pattern

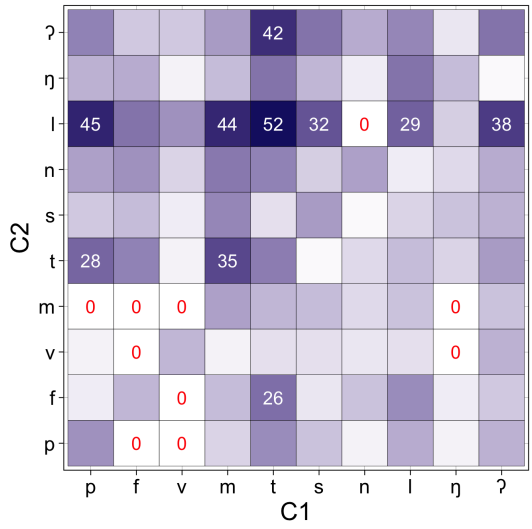
The data I use is taken from Alderete and Bradshaw (2013). Their list contains monomorphemic headwords from the Milner

(1966) dictionary (i.e. unbounded 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.

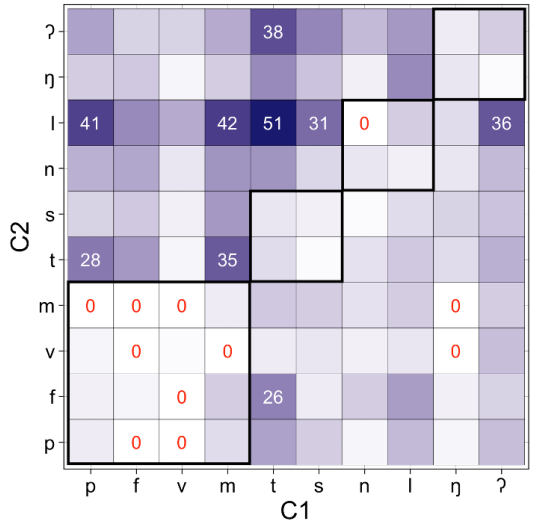
Figures 1 shows counts of all transvocalic consonant-consonant sequences (i.e. C_1VC_2) in these data, where long vowels and diphthongs also count as an intervening V. Fig. 1a uses Alderete and Bradshaw's original root list, while Fig. 1b uses the smaller list with reduplicants removed. C1-C2 combinations 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 and 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 Fig. 1a, 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 Fig. 1b. In general, along the diagonal outlined in Fig. 1b, C1-C2 co-occurrences tend to be less frequent.

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



(a) All roots (n=1640)



(b) No reduplicated forms (n=1498)

Figure 1: Consonant-consonant co-occurrences in Samoan

4.1.2. A statistical model of OCP-place effects

Following Wilson and 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 and 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 consonant 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 (4). 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 Section 6, I also construct a more nuanced phonotactic grammar using the UCLA Phonotactic Learner (Hayes and Wilson, 2008).

(4) *OCP constraints*

Constraint	**Ex. 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

Following Coetzee and Pater (2006), 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 the sound change of *k>ʔ occurred relatively recently, 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 5 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).

Overall, I replicate Alderete and Bradshaw's (2013) findings, and results are consistent with the qualitative observations from Fig. 1. First, the model learned a significant weight for OCP-LAB, showing a strong 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; out of all the constraints that were tested, it was associated with the biggest improvement in log-likelihood. For dorsals, the opposite is true; OCP-BACK is significant, but OCP-BACK-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. Note that existing work on OCP-place disagrees on how much the effect of subsidiary features should be allowed to vary across places of articulation. Coetzee and Pater (2008) allow for the weights of subsidiary features to vary across place, while Wilson and 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.

4.1.3. OCP effects in Proto-Polynesian

Based on a crosslinguistic study of six languages, Krupa (1971) proposes that OCP-place

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 5: OCP constraint weights learned by the phonotactic model

is a general property of Polynesian languages. 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 and 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. 2, 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 therefore where C1-C2 pairs are expected to be under-represented. 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 4.1.2 above. Again, OCP-place constraints were tested for significance using the Likelihood Ratio Test (Hayes et al., 2012). The results, given in Table 6, 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.

4.2. Phonetic similarity and OCP-place

This section presents the results of an acoustic study aimed at quantifying consonant similarity in Samoan. I focus on looking at consonant similarity within the labials and coronals, and do not consider /ŋ/ and /ʔ/. This is because the two sounds form a class of size two, so meaningful comparisons of gradient similarity are not possible. Additionally, /ʔ/ is often elided in natural speech, and is difficult to segment due to its highly variable realization.

As a preview of the results, I find that consonant similarity is a close predictor of gradient OCP-place effects in Samoan, and is consistent with the gradient patterns discussed above in Section 4.1.

In work by Frisch (1996) and Frisch et al. (2004) on OCP-place effects, phonological features are used as a proxy measure of phonetic similarity. Specifically, they quantify the distance between two segments s_1 and s_2 using the equation given in (5).

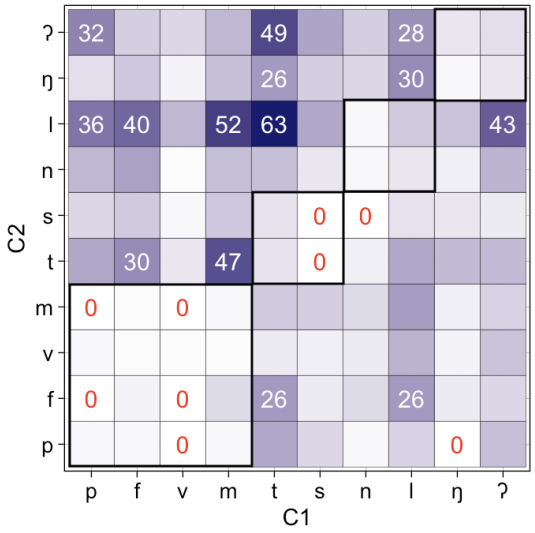


Figure 2: Consonant-consonant co-occurrences in PPn

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 6: OCP constraint weights learned by the phonotactic model for PPn

$$\text{(5)}\quad \text{dist}(s_1,s_2) = \frac{\text{Shared natural classes}}{\text{Shared natural classes} + \text{Non-shared natural classes}}$$

This metric runs into a few potential issues. First, the choice of feature system (and therefore, the resulting natural classes) depends on observations about phonological patterning, and does not necessarily reflect phonetic properties. Feature-based measures also ignore the variable phonetic realization of target phones. In Samoan, for example, [v] is variably lenited and can be closer to sonorants in realization.

Instead, I quantify phonetic similarity as the spectral distance between two phones. This method has been validated in various work on speech sounds, across a range of consonant types (e.g. Gerosa et al., 2006; Mielke, 2012; Cychosz et al., 2019).

4.2.1. Data

Data comes from audio recordings of three male speakers from the Jehovah's Witnesses website.[4] These recordings were done in a quiet setting with minimal to no background noise, and are available in mp3 format (sampling rate: 48 kHz). This corpus faces certain limitations; audio data is only available in compressed format, and is noisier than lab-collected speech.

[4] `https://www.jw.org/en/library/bible/?contentLanguageFilter=sm`

However, it allows for collection of more naturalistic speech, across a variety of contexts and speech rates.

Consonants were manually aligned in Praat TextGrid (Boersma and Weenink, 2023) by a trained phonetician, using visual cues from the waveform and spectrogram. Plosives (all of which are voiceless in Samoan) were marked from onset of the burst to end of aspiration. For consonants where the transition between surrounding vowels is less well-defined, vowel onset/offset was determined by the presence of steady-state formants.

In total, 1866 tokens were aligned and extracted; the distribution of tokens is summarized in Table 7. Note that the tokens are not evenly distributed across phonemes; this reflects the relative token frequency of each phoneme in Samoan.

phone	**N**
p	169
f	242
v	104
m	255
s	183
t	334
l	350
n	229

Table 7: Distribution of extracted tokens

4.2.2. Data analysis and results

Spectral distance was measured by calculating the Euclidean distance between the Mel-frequency cepstral coefficient (MFCC) vectors of two target segments. MFCCs are a small set of features which concisely describe the overall shape of a spectral envelope. They are widely used in speech recognition and have also been applied successfully in phonetics to quantify phonetic distance of phoneme inventories (Mielke, 2012) and coarticulation across a range of consonants varying in place and manner of articulation (Gerosa et al., 2006; Cychosz et al., 2019; Cychosz, 2022).

In general, a greater spectral distance indicates increased acoustic distance between two target sounds. If OCP-place is rooted in similarity avoidance, as proposed by Frisch et al. (2004), we would expect C1-C2 pairs that show strong OCP-place effects to have smaller spectral distance.

To extract MFCCs, the speech signal was first blocked into frames of 15 ms duration, then each speech frame was parameterized into 13 coefficients. Following Gerosa et al. (2006), each MFCC was scaled with the inverse of the standard deviation computed over all data.

Pairwise comparisons of spectral distance were done for every single token. For example, to measure the distance between /p/ and /m/, every token of /p/ was compared against every token of /m/.

4.2.3. Results

Results are summarized in Fig. 3 below; the lefthand figure shows comparisons within labial sounds, while the righthand figure shows comparisons within coronal sounds. The y-axis shows spectral distance, where a larger value indicates that the two segments being compared are acoustically more different.

Overall, results are in line with the phonotactic patterns discussed above First, the spectral distances between labials are lower overall (compared to the distances between coronals), suggesting that they are acoustically more similar to each other. This ties into the phonotactic results, where OCP-LAB was stronger than OCP-COR (and OCP-back).

At the same time, the distances between labials are more compact (there isn't as much variation across different consonant-consonant pairs), and there doesn't appear to be a clear effect of manner. This is consistent with how in the lexical statistics, subsidiary features weren't significant for OCP-LABIAL.

For the coronals, consonant-consonant pairs that mismatch in sonorancy are acoustically

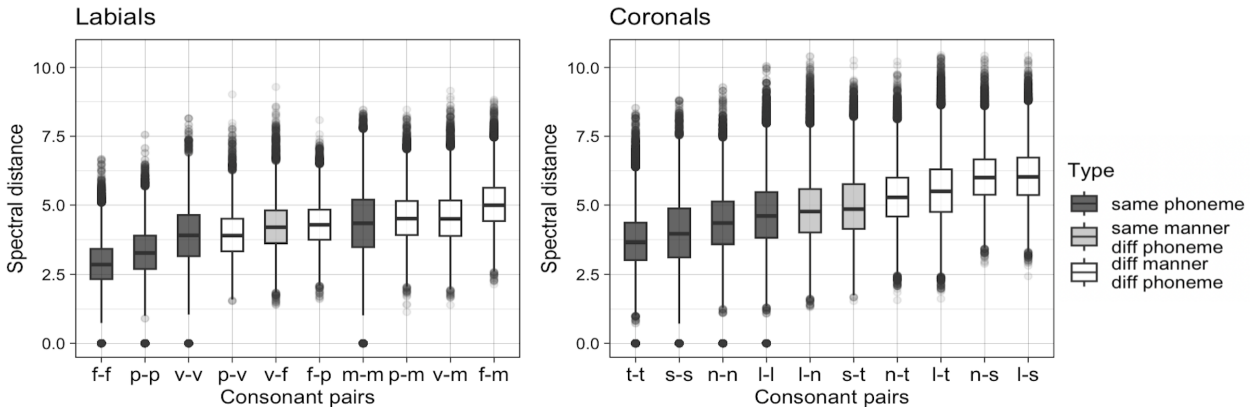


Figure 3: Spectral distances between consonant-consonant pairs

more different (i.e. have a higher spectral distance) than ones that match in sonorancy. In other words, coronal-coronal pairs that match in sonorance (t-t, s-s, n-n, l-n, s-t) are more spectrally similar. Once again, these results are consistent with the phonotactics, where OCP-place effects are stronger for coronals that match in sonorancy.

Overall, the findings of this section suggest that OCP-place in Samoan is closely correlated with the phonetic similarity of the consonants being compared. This supports Frisch's proposal that OCP-place has a phonetic basis and is more concretely rooted in phonetic similarity avoidance.

5. Patterns of reanalysis

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).

POc reconstructions are taken from the Austronesian Comparative Dictionary (ACD; Blust et al., 2023). Protoforms were excluded if they had fewer than six cognates within Oceanic, resulting in a set of 1023 protoforms. Modern Samoan forms are taken from the Milner (1966) dictionary and supplemented with forms from Pratt (1862/1893). As discussed above, I focus on stem-ergative pairs, since it is the most productive of all the suffixes that trigger thematic consonant alternations. The resulting wordlist has 593 stem-suffix pairs.

5.1. 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 8 shows the distribution of final segments in POc and the expected ergative allomorph given this final segment.

Around 68% of stems were either historically vowel-final, or ended in a consonant that was uniformly deleted by regular sound change (and therefore do not reflect as 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/.

Note that in the interest of space, I focus on reanalysis of the consonant-initial allomorphs, and do not consider the relative distribution of /-a/ and /-ina/, or why reanalysis has not been towards the other vowel-initial allomorph /-ia/. Instead, /-a/ and /-ina/ are taken to reflect historically vowel-final allomorphs. Kuo (2023b)

discusses factors that could have affected the distribution of the vowel-initial allormorphs.

POC	Erg	Count	%
*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 8: Distribution of final consonants in POc

Fig. 4 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 omitted. First, we can observe a strong effect of OCP-place for labials; when 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**p**a-**f**ia] or [ta**m**a-**m**ia] are never observed. For coronals and dorsals, the OCP-place effect is weaker. In particular, forms of the type [i**l**a-**n**a], where preceding /l/ is followed by /-na/, are relatively frequent (n=11).

If reanalysis is predictable from statistical distributions within a paradigm, we might expect reanalysis to occur in a way that avoids labial-labial sequences, but not sequences of coronal sonorants.

5.2. Comparing POc and Samoan

In this section, I compare the distributional patterns of thematic consonants 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.

Fig. 5 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 underestimate 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 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.

Fig. 6 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 still true in 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 *bu**liŋ** (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. This process was repeated 10,000 times to pro-

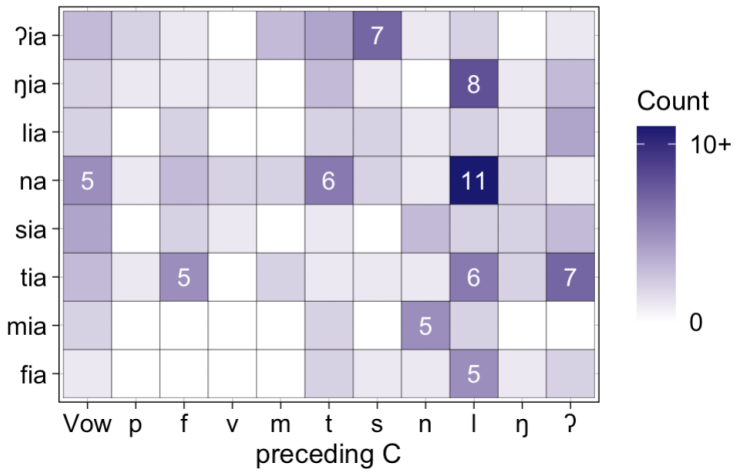


Figure 4: Expected distribution of ergative allomorphs by identity of preceding consonant (POc)

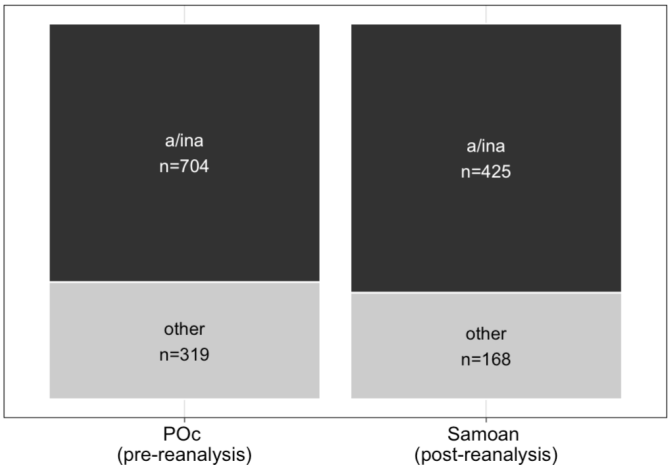


Figure 5: Distribution of ergative allomorphs before and after reanalysis

duce 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. 7 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. 8 visualizes the Monte Carlo results for the subset of C_{prev}-C_{theme} pairs where at

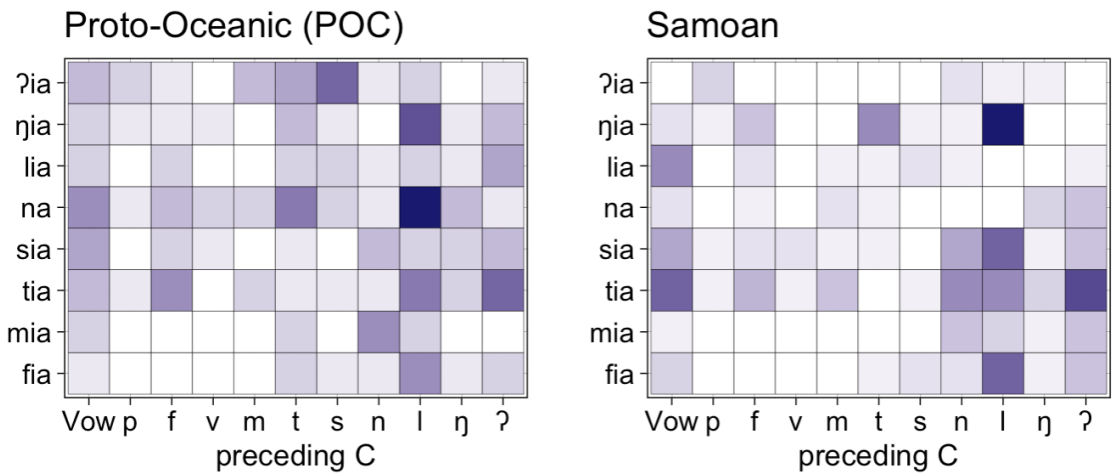


Figure 6: Distribution of allomorphs by preceding consonant in POc vs. Samoan

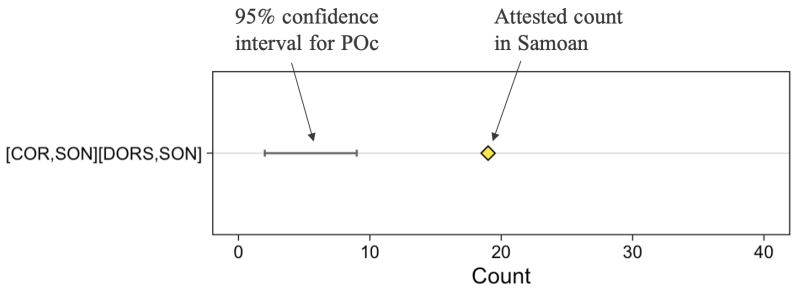


Figure 7: Chance-level distribution of C_{prev}-C_{theme} pairs vs. observed count in Samoan

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. [LAB,OBS][COR,SON] pairs (e.g. [ipo-lia], [ifo-na]) are slightly underattested. Most strikingly, however, [COR,SON][COR,SON] pairs are highly under-attested. In other words, the number of suffixed forms like [ilo-na] is lower in Samoan than would be expected given the POc distribution.

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

5.3. Direct evidence of reanalyses

In this section, I consider the subset of stem-ergative pairs which have known POc proto-forms (n=147); for the forms, the exact directions of reanalysis are known. Table 9 summarizes the proportion of forms that have undergone reanalysis based on what allomorph they would have taken historically. Fig. 9 visualizes this same data. Overall, results are consistent with the conclusions of the previous section.

First, where reanalysis has occurred, it is most often towards /-a/ and /-ina/ (labeled here as /-(in)a/), rather than another allo-

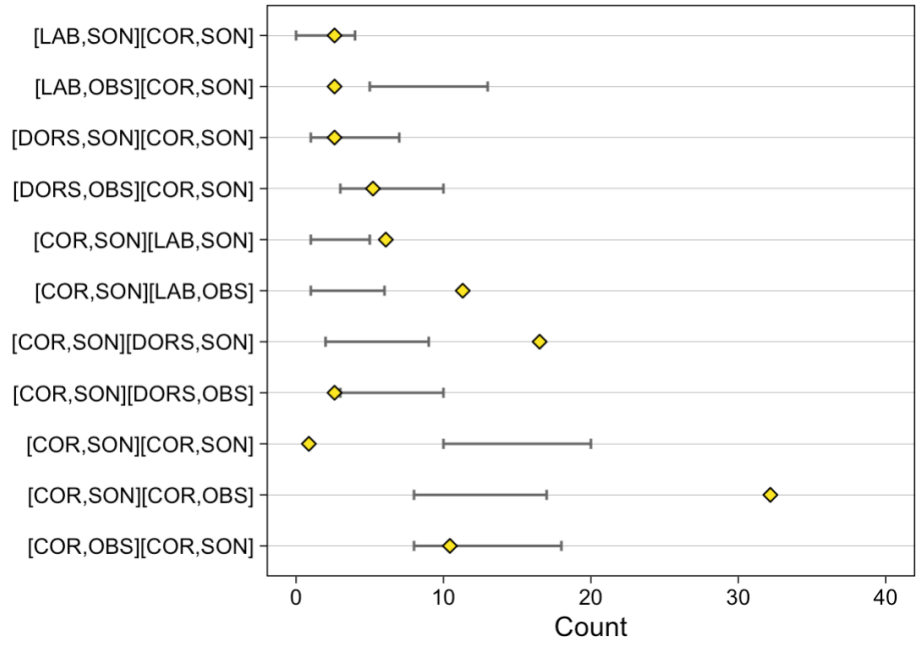


Figure 8: Chance-level distribution of C_{prev}-C_{theme} pairs vs. observed count in Samoan (the most frequent reflex is shown in bold)

morph. Samoan stems that are expected to take /(in)a/ have also undergone the least amount of reanalysis; just 8% of these have been reanalyzed. In contrast, for stems that historically took a /-Cia/ allomorph, there has been more extensive reanalysis.

Table 10 breaks down reanalysis involving /-lia/ and /-na/ forms by whether the stem has an immediately preceding /l/ or /n/. 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.

5.4. Interim summary

Overall, the picture that emerges is that a mismatch between stem phonotactics and morphophonological alternations was gradually removed over time. In other words, I propose 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.

6. Modeling reanalysis

In this section, I test whether reanalysis of Samoan thematic consonants can be modeled as the combined effects of frequency and markedness.

The model has three main components. First, it uses Maximum Entropy Harmonic Grammar (MaxEnt; Smolensky, 1986; Goldwater and Johnson, 2003), a probabilistic variant of Optimality Theory. Additionally, markedness bias is implemented in the model as a Gaussian prior, following the methodology of

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

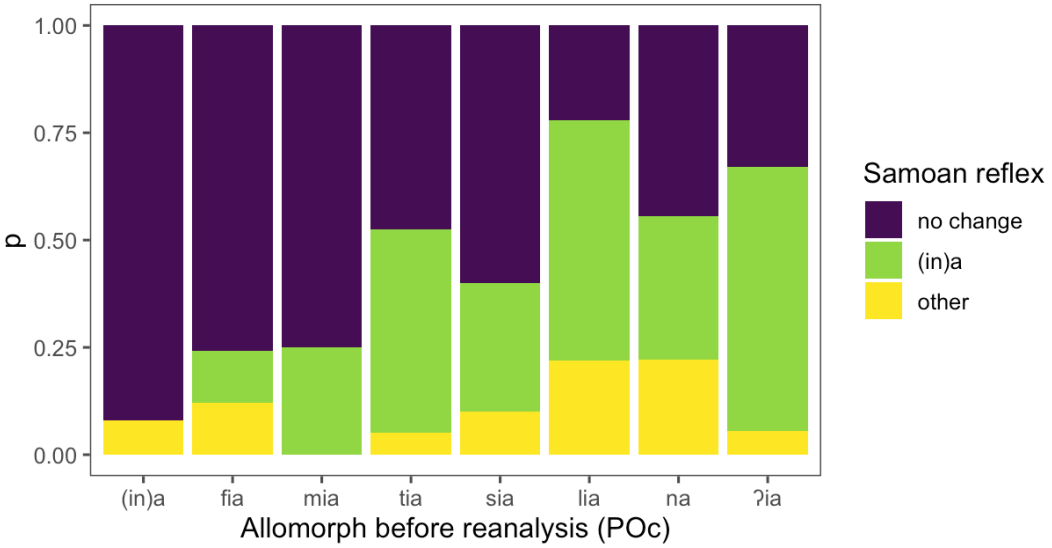


Figure 9: Summary of reanalyses (POc protoforms vs. Samoan reflexes)

Wilson (2006) and White (2013, 2017). This biased model will be compared to control models that do not have a markedness bias. Finally, to mirror the cumulative effect of reanalyses over time, the model has an iterative component, in which the output of one iteration of the model becomes the input for the next. In the interest of space, I do not go over the model in detail; the reader is referred to Kuo (2023b) for the specifics of model implementation.

The rest of this section is organized as follows. Section 6.1 will discuss the details of the MaxEnt model implementation. Sections 6.2 and 6.3 introduce my method for incorporating a markedness bias rooted in phonotactics. Following this, Section 6.4 describes the iter-

POc	has /n,l/	Samoan	N	Example
lia	yes	lia	0	[ana-lia] (<*anal)[5]
	yes	other	4	[fono-tia] (<*ponor, cf. *[fono-lia])
	no	lia	2	[tautau-lia] (<*saur)
	no	other	3	[fata-ina] (*pantar, cf. *[fata-lia])
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 10: Reanalyses of /lia/ and /na/

ated learning component. Finally, Section 6.5 presents modeling results.

6.1. 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 (11), 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 Table 11, makes morphophonology more complex. On the other hand, stem URs are closer to their surface forms.

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, 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 Table 12 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.

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 (6) for the stem [tuʔu].

(6) *[tuʔu] with different ergative allomorphs*

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

Note that although I adopt the morphological approach, both the phonological and morphological approaches can be used to model reanalysis. In Kuo (2023b), a model implementation which adopts the phonological approach is explored in more detail.

An example of a model input and its corresponding candidates is given in (7). 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.

		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/

Table 11: URs under phonological vs. morphological analysis

STEM	SUFFIXED FORMS	GLOSS
alofa	alofa**ŋ**ia, alofa**ʔ**aŋa	'love, affection'
eʔe	eʔe**t**ia, eqe**n**aʔi	'be propped up/raised'
au	au**l**ia, au**∅**aʔi	'flow on, continue
tae	taːe**∅**a, tae**n**aʔi	'gather'

Table 12: Variation in thematic consonant across suffixal contexts

(7) *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. The model inputs reflect several simplifying assumptions. For the suffixed form candidates, /-ina/ and /-a/ are combined, since their relative distribution is not the focus of the current paper. 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/.

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, 1995; Kager, 1996).

This is shown in the tableau in (8), 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 allomorph. Consequently, candidate (a) has the highest predicted probability. Exponence constraints can also interact with markedness constraints. Here, OCP-COR-SON penalizes candidate (b), causing it to be assigned the lowest predicted probability.

(8) *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

6.2. Implementing a soft markedness bias

In MaxEnt, bias can be implemented by giving certain constraints a preference 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 (9), 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.

$$(9) \quad \sum_{i=1}^{m} \frac{(w_i - \mu_i)^2}{2\sigma^2}$$

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 a fixed value of 1.5 for all constraints. μ is varied to implement different learning biases. More concretely, constraint(s) which enforce OCP-place, to be discussed below, are given higher μ values than competing morpheme exponence constraints.

6.3. Deriving markedness effects from phonotactics

In Section 4.1, I argue that OCP-place effects can be derived directly from stem phonotactics. To test this, I propose the following procedure to derive markedness directly from phonotactics, schematized in Fig. 10.

First, phonotactic grammars are trained on monomorphemic roots using the UCLA Phonotactic Learner (UCLAPL; Hayes and 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 using the procedure described above.

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 and Clark, 2011); this is the same corpus used earlier in Section 4.1. As described there, this corpus had polymorphemic items removed, and was modified to reflect the regular sound changes that have happened between PPn and Samoan.

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 three phonotactic models using the same corpus of 1645 roots used in Section 4.1[6]:

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, consist-

[6] A fourth model, which includes constraints against vowel hiatus, is discussed in Kuo (2023b)

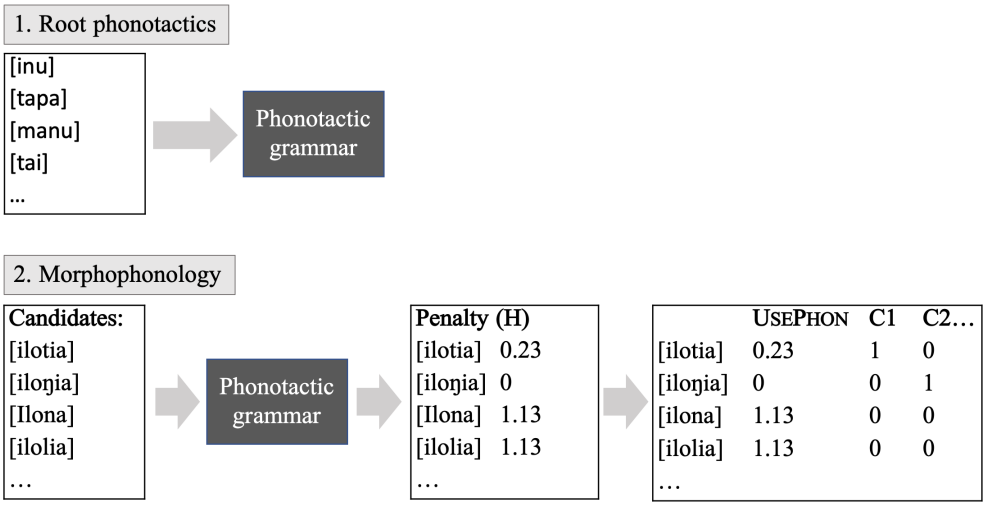


Figure 10: Incorporating phonotactic markedness into morphophonological grammar

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

The three models are summarized in (10) 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 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.

(10) *Phonotactic grammar parameters*

Model	**Parameters**
NATURAL CLASS	50
OCP	20
BIGRAM	100

These models are compared to help us gain insight into how speakers' phonotactic knowledge can affect reanalysis. The NATURAL CLASS and OCP 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 models perform better, this suggests that speakers prefer to generalize patterns to natural classes. The OCP model is additionally more restrictive, in that they only allows for phonetically motivated OCP constraints, rather than potentially arbitrary constraints learned over any natural class. If the OCP model outperforms 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.

6.4. Iterated learning

To simulate reanalysis over time, I use adopted an iterated learning paradigm, where one iteration of the model becomes the input to the next iteration. There are various ap-

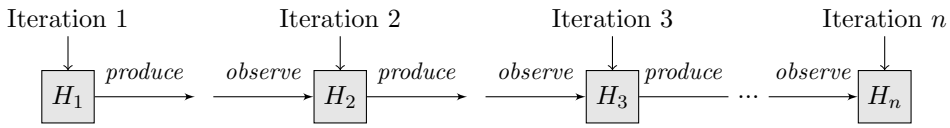


Figure 11: Structure of an iterated learning model, adapted from Ito and Feldman (2022, p. 3). H_i indicates hypotheses of each generation.

proaches to modeling generational learning, including phonological rules that apply variably (Weinreich et al., 1968), dynamical systems (Niyogi, 2006), connectionist frameworks (Tabor, 1994), competing grammars (Yang, 1976), exemplar-based frameworks (Pierrehumbert, 2002), and more recently variants of Optimality Theory (e.g. Boersma, 1998; Zuraw, 2003).

I assume a relatively simple 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.

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 iteration has just one agent and one learner, as illustrated in Fig. 11. In the first iteration, the agent A1 produces the output language based on their grammatical knowledge (i.e. Hypothesis 1; H_1). A hypothesis is essentially 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 iteration, not all of the information of the language is presented, resulting in a learning "bottleneck" (Brighton, 2002; Kirby, 2001; Griffiths and Kalish, 2007). This bottleneck causes patterns that are easier to learn to be preferentially passed down to the next iteration, and become more prominent over time. In the current study, I follow Ito and Feldman (2022), and implement the bottleneck 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.).

Note that the iterated learning paradigm I adopt doesn't consider the interaction of multiple Agents, 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. In fact, Baker (2008) finds that such multi-agent models produce more empirically accurate results.

The iterated 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. I test 5 forgetting rates (0.05, 0.1, 0.15, 0.2, 0.25). In the interest of brevity, and because the model trended in the same direction across all five forgetting rates, the rest of this paper will only present models with a forgetting rate of 0.2. The number of iterations is set to 30.

Because random sampling causes each iteration of the model to vary slightly, all subsequent models were run 30 times, and predicted probability values are the mean of these 30 trials.

6.5. Model specifications and results

In the markedness-biased models, the phonotactically motivated UsePhonotactics constraint is given a bias towards higher weight;

specifically, it is assigned a higher μ value of 3, while μ = 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 13 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**
BIGRAM	-2438.39	10.42

Table 13: Model results: log likelihood

Of the markedness-biased models, the BIGRAM model performs the worst; this suggests that models which generalize to natural classes are better predictors of learner behavior.

Of the other two models, the OCP grammar does better than the NATURAL CLASS 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.

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

CONSTRAINT	w	PENALIZES…
*[n...{l,s}]	1.26	ino-lia, ino-sia
*[{l,n}...l]	0.93	ino-lia, ilo-lia
*[l...n]	0.78	ilo-na

Based on the POc data, stems of the type [i**n**o-**n**a], [i**no**-**l**ia], and [i**l**o-**l**ia] are expected to be infrequent, but [i**l**o-**n**a] stems were relatively frequent. Because of this, the NATURAL CLASS grammar does not learn a general OCP-COR-SON constraint, but instead learns the three separate constraints given in (11). The constraint *[l...n] is assigned a relatively lower weight, so the model does not penalize [ilo-na] type words as heavily and underpredicts 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.

Fig. 12 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).

7. Conclusion

In this paper, I argue for a model of reanalysis where frequency-matching is modulated by a markedness bias. In particular, I find that reanalysis of the ergative suffix is generally towards the more frequent allomorphs, but also happens in a way that avoids violations of OCP-place. These results are confirmed with a model of reanalysis that is based in MaxEnt, with the markedness bias implemented as a Gaussian prior.

Results of this paper also support the idea that markedness bias is constrained in various ways. First, OCP-place effects are active in Samoan stem phonotactics, in line with the active markedness restriction. Additionally, an acoustic study of consonant similarity suggests that these OCP-place effects are The Samoan

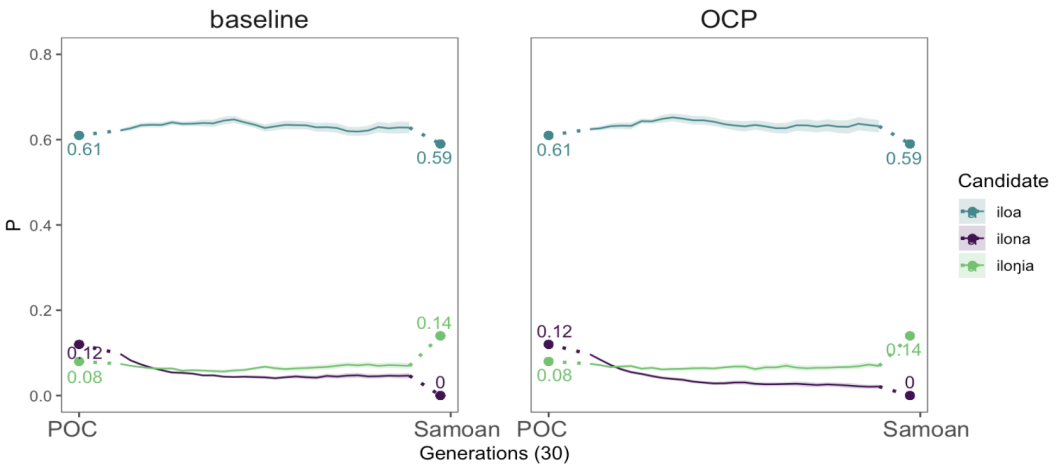


Figure 12: Model predictions in stems with a preceding /l/

data, in particular, is compatible with idea that markedness effects must already be active in the language's stem phonotactics, and be rooted in phonetic naturalness.

In Section 6, I also compared between three phonotactic models. All three models were 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 is the most restrictive, 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.

Notably, as Glewwe (2019) points out, markedness effects of the variety discussed in this paper 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 and 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.

Finally, although it is not the focus of this paper, my results also provide evidence for Frisch et al.'s (2001) proposal that phonotactic regularities have a functional diachronic origin. Their proposal for Arabic OCP-place effects is that a processing constraint against sequences of similar sounds led to changes that removed sequences of homorganic consonants. This resulted in the synchronic phonotactic pattern where OCP-place is strongly present. In my acoustic study, I find similar support that OCP-place is rooted in phonetic similarity avoidance. Notably, this view contrasts with McCarthy's (1988; 1994) classical OT analysis of OCP-place in Arabic, where constraints are selected from a universal inventory of possible constraints.

References

Albright, A., 2002a. A restricted model of UR discovery: Evidence from Lakhota. Ms, University of California at Santa Cruz .

Albright, A., 2010. Base-driven leveling in Yiddish verb paradigms. NLLT 28, 475–537.

Albright, A., Hayes, B., 2003. Rules vs. analogy in English past tenses: A computational/experimental study. Cognition 90, 119–161.

Albright, A.C., 2002b. The identification of bases in morphological paradigms. Ph.D. thesis. University of California, Los Angeles.

Alderete, J., Bradshaw, M., 2013. Samoan root phonotactics: Digging deeper into the data. Linguistic Discovery 11.

Bailey, T.M., Hahn, U., 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? Journal of Memory and Language 44, 568–591.

Baker, A., 2008. Computational approaches to the study of language change. Language and Linguistics Compass 2, 289–307.

Becker, M., Ketrez, N., Nevins, A., 2011. The surfeit of the stimulus: Analytic biases filter lexical statistics in Turkish laryngeal alternations. Language , 84–125.

Berkley, D.M., 1994. The OCP and gradient data. Studies in the Linguistic Sciences 1/2, 59–72.

Berkley, D.M., 2000a. Gradient obligatory contour principle effects. Ph.D. thesis. Northwestern University.

Berkley, D.M., 2000b. Gradient OCP Effects. Ph.D. thesis. Northwestern University.

Blevins, J., 1994. A phonological and morphological reanalysis of the Maori passive. Te Reo 37, 29–53.

Blevins, J., 2004. Evolutionary phonology: The emergence of sound patterns. Cambridge University Press.

Blust, R., Trussel, S., Smith, A.D., 2023. CLDF dataset derived from Blust's "Austronesian Comparative Dictionary" (v1.2) [data set]. Zenodo. URL: https://doi.org/10.5281/zenodo.7741197.

Boersma, P., 1998. Functional Phonology: Formalizing the interactions between articulatory and perceptual drives. The Hague: Holland Academic Graphics.

Boersma, P., Weenink, D., 2023. Praat (version 6.3.17) [software]. Latest version available for download from www.praat.org.

Brighton, H., 2002. Compositional syntax from cultural transmission. Artificial life 8, 25–54.

Buckley, E., 1997. Tigrinya root consonants and the OCP. University of Pennsylvania Working Papers in Linguistics 4, 3.

Chomsky, N., Halle, M., 1968. The sound pattern of English. ERIC.

Chong, A.J., 2019. Exceptionality and derived environment effects: a comparison of Korean and Turkish. Phonology 36, 543–572.

Chong, A.J., 2021. The effect of phonotactics on alternation learning. Language 97, 213–244.

Chung, S., 1978. Case marking and grammatical relations in Polynesian. Ph.D. thesis. Harvard University.

Churchward, S., 1951. A Samoan grammar. Spectator Publishing Company.

Clark, D.R., 1973. Aspects of proto-Polynesian syntax. Ph.D. thesis. University of California, San Diego.

Coetzee, A.W., Pater, J., 2006. Lexically ranked OCP-Place constraints in Muna. Ms, University of Michigan and University of Massachusetts, Amherst.

Coetzee, A.W., Pater, J., 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. NLLT 26, 289–337.

Coleman, J., Pierrehumbert, J., 1997. Stochastic phonological grammars and acceptability, in: 3rd meeting of the ACL Special Interest Group in computational phonology: Proceedings of the workshop. Association for Computational Linguistics, pp. 49–56.

Cook, K.W., 1988. A cognitive analysis of grammatical relations, case, and transitivity in Samoan. University of California, San Diego.

Cychosz, M., 2022. Language exposure predicts children's phonetic patterning: Evidence from language shift. Language 98, 461–509. Publisher: NIH Public Access.

Cychosz, M., Edwards, J.R., Munson, B., Johnson, K., 2019. Spectral and temporal measures of coarticulation in child speech. The Journal of the Acoustical Society of America 146, EL516–EL522. Publisher: Acoustical Society of America.

Daelemans, W., Zavrel, J., Van Der Sloot, K., Van den Bosch, A., 2004. Timbl: Tilburg memory-based learner. Tilburg University .

Daugherty, K.G., Seidenberg, M.S., 1994. Beyond rules and exceptions, in: Lima, S.D., Corrigan, R., Iverson, G.K. (Eds.), The reality of linguistic rules. John Benjamins Publishing, pp. 353–388.

Dell, G.S., 1984. Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. Journal of Experimental Psychology: Learning, Memory, and Cognition 10, 222–233. URL: http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.10.2.222, doi:10.1037/0278-7393.10.2.222.

Dempwolff, O., 1929. Das austronesische Sprachgut in den polynesischen Sprachen. G. Kolff.

Eberhard, D.M., Simons, G.F., (eds), C.D.F., 2023. Ethnologue: Languages of the World (26th edition). Dallas, Texas: SIL International. URL: http://www.ethnologue.com.

Eddington, D., 1996. Diphthongization in Spanish derivational morphology: An empirical investigation. Hispanic Linguistics 8, 1–13.

Eddington, D., 1998. Spanish diphthongization as a non-derivational phenomenon. Rivista di Linguistica 10, 335–354.

Eddington, D., 2004. Spanish Phonology and Morphology: Experimental and Quantitative Perspectives. John Benjamins Publishing Company.

Ernestus, M.T.C., Baayen, R.H., 2003. Predicting the unpredictable: Interpreting neutralized segments in Dutch. Language 79, 5–38.

Frisch, S., 1996. Similarity and frequency in phonology. Ph.D. thesis. Northwestern University.

Frisch, S.A., Pierrehumbert, J.B., Broe, M.B., 2004. Similarity avoidance and the OCP. Language & Linguistic Theory 22, 179–228.

Frisch, S.A., Zawaydeh, B.A., 2001. The psychological reality of OCP-Place in Arabic. Language 77, 91–106.

Gallagher, G., Coon, J., 2009. Distinguishing total and partial identity: Evidence from Chol. NLLT 27, 545–582.

Gerosa, M., Lee, S., Giuliani, D., Narayanan, S., 2006. Analyzing children's speech: An acoustic study of consonants and consonant-vowel transition, in: IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, pp. 393–96.

Glewwe, E.R., 2019. Bias in phonotactic learning: Experimental studies of phonotactic implicationals. University of California, Los Angeles.

Goldwater, S., Johnson, M., 2003. Learning OT constraint rankings using a maximum entropy model, in: Proceedings of the Stockholm workshop on variation within Optimality Theory, pp. 111–120.

Greenberg, J., 1950. The patterning of root morphemes in Semitic. Word 6, 162–181.

Greenhill, S.J., Clark, R., 2011. POLLEX-online: The Polynesian lexicon project online. Oceanic Linguistics , 551–559.

Griffiths, T.L., Kalish, M.L., 2007. A Bayesian view of language evolution by iterated learning. Cognitive Science 31, 441–480.

Hale, K., 1968. Review of Hohepa (1967)—'a profile generative grammar of Maori'. Journal of the Polynesian Society 77, 83–99.

Hale, K., 1973. Deep-surface canonical disparities in relation to analysis and change: An Australian example, in: Sebeok, T. (Ed.), Current Trends in Linguistics. The Hague: Mouton. volume 11, pp. 401–458.

Hale, M., Reiss, C., 2000. "Substance abuse" and "dysfunctionalism": Current trends in phonology. Linguistic Inquiry 31, 157–169.

Hare, M., Elman, J.L., 1995. Learning and morphological change. Cognition 56, 61–98.

Hayes, B., 2004. Phonological acquisition in optimality theory: the early stages, in: Kager, R., Pater, J., Zonneveld, W. (Eds.), Constraints in phonological acquisition. Cambridge University Press, pp. 158–203.

Hayes, B., Jo, J., 2020. Balinese stem phonotactics and the subregularity hypothesis. Ms, UCLA.

Hayes, B., Londe, Z.C., 2006. Stochastic phonological knowledge: The case of Hungarian vowel harmony. Phonology 23, 59–104.

Hayes, B., Siptár, P., Zuraw, K., Londe, Z., 2009. Natural and unnatural constraints in Hungarian vowel harmony. Language , 822–863.

Hayes, B., Wilson, C., 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry 39, 379–440.

Hayes, B., Wilson, C., Shisko, A., 2012. Maxent grammars for the metrics of Shakespeare and Milton. Language 88, 691–731.

Hovdhaugen, E., et al., 1986. The chronology of three Samoan sound changes, in: Focal II: Papers from the fourth international conference on Austronesian linguistics, Pacific Linguistics. pp. 313–331.

Hudson Kam, C.L., Newport, E.L., 2005. Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. Language learning and development 1, 151–195.

Hudson Kam, C.L., Newport, E.L., 2009. Getting it right by getting it wrong: When learners change languages. Cognitive psychology 59, 30–66.

Ito, C., Feldman, N.H., 2022. Iterated learning models of language change: A case study of sino-korean accent. Cognitive Science 46, e13115.

Jarosz, G., 2006. Rich lexicons and restrictive grammars: maximum likelihood learning in Optimality Theory. Ph.D. thesis. Johns Hopkins University.

Jun, J., Lee, J., 2007. Multiple stem-final variants in Korean native nouns and loanwords. Journal of the Linguistic Society of Korea 47, 159–187.

Kager, R., 1996. On affix allomorphy and syllable counting, in: Ursula, K. (Ed.), Interfaces in phonology. Akademie Verlag, pp. 155–171.

Kang, Y., 2006. Neutralizations and variations in Korean verbal paradigms. Harvard Studies in Korean Linguistics 11, 183–196.

Kawahara, S., Ono, H., Sudo, K., 2006. Consonant co-occurrence restrictions in Yamato Japanese. Japanese/Korean Linguistics 14, 27–38.

Kenstowicz, M., 1996. Base-identity and uniform exponence: alternatives to cyclicity, in: Durand, J., Laks, B. (Eds.), Current Trends in Phonology: Models and Methods. Salford: University of Salford, pp. 363–394.

Kiparsky, P., 1965. Phonological change. Ph.D. thesis. Massachusetts Institute of Technology.

Kiparsky, P., 1978. Analogical change as a problem for linguistic theory. Studies in the Linguistic Sciences Urbana, Ill 8, 77–96.

Kiparsky, P., 1997. Covert generalization, in: Mediterranean Morphology Meetings, pp. 65–76.

Kirby, S., 2001. Spontaneous evolution of linguistic structure-an iterated learning model of the emergence of regularity and irregularity. IEEE Transactions on Evolutionary Computation 5, 102–110.

Krupa, V., 1966. The phonemic structure of bi-vocalic morphemic forms in Oceanic languages. The Journal of the Polynesian Society 75, 458–497.

Krupa, V., 1967. On phonemic structure of morpheme in Samoan and Tongan. Beitrèage zur Linguistik und Informationsverarbeitung 12, 72–83.

Krupa, V., 1971. The phonotactic structure of the morph in Polynesian languages. Language 47, 668–

684.

Kuo, J., 2023a. Evidence for prosodic correspondence in the vowel alternations of tgdaya seediq. Phonological Data and Analysis 5, 1–31.

Kuo, J., 2023b. Phonological markedness effects in reanalysis. Ph.D. thesis. University of California, Los Angeles.

Labov, W., 1994. Principles of Linguistic Change, Volume 1: Internal Factors. Wiley-Blackwell.

de Lacy, P., 2002. Maximal words and the Maori passive, in: Richards, N. (Ed.), Proceedings of the Austronesian Formal Linguistics Association (AFLA) 8. MIT Working Papers in Linguistics.

de Lacy, P., 2003. Maximal words and the Maori passive, in: McCarthy, J. (Ed.), Optimality Theory in phonology: A reader. Blackwell, pp. 495–512.

Lichtenberk, F., 2001. On the morphological status of thematic consonants in two Oceanic languages, in: Bradshaw, J., Rehg, K.L. (Eds.), Issues in Austronesian Morphology: A festschrift for Byron W. Bender. Pacific Linguistics, pp. 123–147.

Ling, C., Marinov, M., 1993. Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. Cognition 49, 235–290.

MacWhinney, B., Leinbach, J., 1991. Implementations are not conceptualizations: Revising the verb learning model. Cognition 40, 121–157.

Martin, A., 2011. Grammars leak: Modeling how phonotactic generalizations interact within the grammar. Language 87, 751–770. doi:10.1353/lan.2011.0096.

McCarthy, J.J., 1988. Feature geometry and dependency: A review. Phonetica 45, 84–108.

McCarthy, J.J., 1994. The phonetics and phonology of Semitic pharyngeals, in: Keating, P. (Ed.), Phonological structure and phonetic form. Cambridge University Press, pp. 191–233.

Mielke, J., 2012. A phonetically based metric of sound similarity. Lingua 122, 145–163. URL: https://linkinghub.elsevier.com/retrieve/pii/S0024384111000891, doi:10.1016/j.lingua.2011.04.006.

Milner, G.B., 1966. Samoan Dictionary; Samoan-English, English-Samoan. ERIC.

Moore-Cantwell, C., 2008. Samoan thematic consonants and the -Cia suffix.

Moreton, E., Pater, J., 2012a. Structure and substance in artificial-phonology learning, part I: Structure. Language and linguistics compass 6, 686–701.

Moreton, E., Pater, J., 2012b. Structure and substance in artificial-phonology learning, part {II}: Substance. Language and linguistics compass 6, 702–718.

Mosel, U., Hovdhaugen, E., 1992. Samoan reference grammar. Scandinavian Univ. Press.

Niyogi, P., 2006. The computational nature of language learning and evolution. MIT press Cambridge, MA.

Nosofsky, R.M., 1990. Relations between exemplar-similarity and likelihood models of classification. Journal of Mathematical psychology 34, 393–418.

Nosofsky, R.M., 2011. The generalized context model: An exemplar model of classification, in: Pothos, E.M., Wills, A.J. (Eds.), Formal approaches in categorization. Cambridge University Press, pp. 18–39.

Oh, Y., Todd, S., Beckner, C., Hay, J., King, J., Needle, J., 2020. Non-Māori-speaking New Zealanders have a Māori proto-lexicon. Scientific reports 10, 1–9.

Ohala, J.J., 1993. Sound change as nature's speech perception experiment. Speech Communication 13, 155–161.

Padgett, J., 1991. Stricture in Feature Geometry. Ph.D. thesis. University of Massachusetts, Amherst.

Padgett, J., 1995. Stricture in Feature Geometry. Dissertations in Linguistics. CSLI Publications.

Pater, J., Tessier, A.M., 2005. Phonotactics and alternations: Testing the connection with artificial language learning. University of Massachusetts Occasional Papers in Linguisitcs 31, 1–16.

Pawley, A., 1962. The person-markers in Samoan. Te Reo 5, 52–56.

Pawley, A., 1966. Polynesian languages: A subgrouping based on shared innovations in morphology. The Journal of the Polynesian Society 75, 39–64.

Pawley, A., 1967. The relationships of Polynesian Outlier languages. The Journal of the Polynesian Society 76, 259–296.

Pawley, A., 2001. Proto polynesian *-CIA, in: Bradshaw, J., Rehg, K.L. (Eds.), Issues in Austronesian Morphology: A festschrift for Byron W. Bender. Pacific Linguistics, pp. 193–216.

Pierrehumbert, J., 1993. Dissimilarity in the Arabic verbal roots, in: Proceedings of the Northeast Linguistic Society, University of Massachusetts Amherst. pp. 367–381.

Pierrehumbert, J., 2002. Word-specific phonetics, in: Gussenhoven, C., Warner, N. (Eds.), Laboratory phonology VII. Berlin: Mouton de Gruyter, pp. 101—140.

Pierrehumbert, J.B., 2006. The statistical basis of an unnatural alternation. Laboratory phonology 8, 81–107.

Pratt, G., 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.

Rumelhart, D.E., McClelland, J.L., 1987. Learning the past tenses of English verbs: Implicit rules or parallel distributed processing?, in: MacWhinney, B. (Ed.), Mechanisms of language aquisition. Lawrence Erlbaum Associates, Inc, pp. 195—-248.

Russell, K., 1995. Morphemes and candidates in Optimality Theory. Rutgers Optimality Archive 44.

Sanders, G., 1990. On the analysis and implications of Maori verb alternations. Lingua 80, 149–196. doi:10.1016/0024-3841(90)90019-H.

Schumacher, R.A., Pierrehumbert, J.B., 2021. Familiarity, consistency, and systematizing in morphology. Cognition 212, 104512.

Sevald, C.A., Dell, G.S., 1994. The sequential cuing effect in speech production. Cognition 53, 91–127. URL: https://linkinghub.elsevier.com/retrieve/pii/0010027794900671, doi:10.1016/0010-0277(94)90067-1.

Skousen, R., 1989. Analogical Modeling of Language. Springer Netherlands.

Smolensky, P., 1986. Information processing in dynamical systems: Foundations of harmony theory. Technical Report. Colorado Univ at Boulder Dept of Computer Science.

Steriade, D., 2001. The phonology of perceptibility effects: The p-map and its consequences for constraint organization.

Tabor, W., 1994. Syntactic innovation: A connectionist model. Ph.D. thesis. Stanford University.

Tesar, B., Prince, A., 2003. Using phonotactics to learn phonological alternations. CLS 39, 241–269.

Violette, P.L., 1880. Dictionnaire samoa-français-anglais. Maisonneuve.

Weinreich, U., Labov, W., Herzog, M., 1968. Empirical foundations for a theory of language change. University of Texas Press.

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

White, J.C., 2013. Bias in phonological learning: Evidence from saltation. Ph.D. thesis. UCLA.

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

Wilson, C., Obdeyn, M., 2009. Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms, Johns Hopkins University.

Yang, C., 2016. The price of linguistic productivity: How children learn to break the rules of language. MIT press.

Yang, H.f., 1976. The phonological structure of the paran dialect of Sediq. Bulletin of the Institute of History and Philology Academia Sinica 47, 611–706.

Yip, M., 1989. Feature geometry and co-occurrence restrictions. Phonology 6, 349–374.

Zamuner, T.S., 2006. Sensitivity to word-final phonotactics in 9-to 16-month-old infants. Infancy 10, 77–95.

Zuraw, K., 2002. Aggressive reduplication. Phonology 19, 395–439.

Zuraw, K., 2003. Probability in language change, in: Bod, R., Hay, J., Jannedy, S. (Eds.), Probabilistic Linguistics. MIT Press, pp. 139–176.

Zuraw, K., Kristine, M.Y., Orfitelli, R., 2014. The word-level prosody of Samoan. Phonology 31, 271–327.

Zuraw, K.R., 2000. Patterned exceptions in phonology. Ph.D. thesis. University of California, Los Angeles.