Phonetic naturalness in the reanalysis of Samoan thematic consonant alternations

Jennifer Kuo

[a] *Cornell University, Department 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 in addition, reanalysis may be constrained by phonotactics (global distributional regularities in the lexicon) and phonetic naturalness. More concretely, I find that reanalysis of Samoan thematic consonants generally matches distributional patterns within the paradigm. However, reanalysis is also modulated by a phonotactic dispreference against sequences of homorganic consonants, analyzed here in Optimality Theoretic terms by OCP-place. These results are confirmed in an iterated learning model that is based in MaxEnt (Goldwater and Johnson, 2003). 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, substantive bias, phonotactics, OCP-place, phonetic similarity

1. Introduction

It is well established that phonetic detail can lead to sound change, and phonetic variation can evolve into phonological processes (phonologization; Hyman, 1976; Ohala, 1993; Ramsammy, 2015). However, the effect of phonetic detail on the restructuring of paradigmatic alternations is less well-established. In this paper, I address this issue, and focus on how changes to morphophonological paradigms are constrained by phonetic naturalness. In addition to looking at patterns of paradigm restructuring, I present an acoustic study that quantifies the phonetic similarity between segments; these results are consistent with the proposal that changes to paradigms are sensitive to phonetic detail.

Additionally, 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. As such, findings from this paper not only improve our understanding of diachrony, but also have the potential to provide insight into the role of phonetic detail in morphophonological learning.

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.

Output	**Real-word examples**
gleeded	*want, need, start, decide*
gled	*read, lead, bleed, breed*
glode	*speak, freeze, weave*
gleed	*shed, spread, put*

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

Ambiguity can in turn result in acquisition errors (e.g. *go/goed* instead of *go/went* in English). When such errors are adopted into the speech community, they result in a type of change over time I refer to as *reanalysis*. Some examples of reanalysis in English past tense include *help/halp*→*help/helped* (∼1300, OED) and *dive/dived*→*dive/dove* (∼1800, OED).[1]

In this paper, I investigate what factors learners are sensitive to when deciding the direction of reanalysis. Existing models of reanalysis argue that learners are sensitive to probabilistic distributions *within* the paradigm. I propose that in addition to this *local* distributional information, learners are also able to draw on phonotactics, or *global* distributional information about how segments can be combined in the language. Specifically, learners appear to selectively utilize phonotactics that are rooted in phonetic naturalness.

The empirical focus of my paper is reanalysis involving Samoan verbal paradigms. I present data suggesting that while reanalysis is generally in the direction predicted by local distributional information, it is also modulated by effects of a phonotactic generalization that is rooted in phonetic naturalness. These results are confirmed using iterated learning models that simulate the cumulative effect of reanalyses over time. In particular, a model that uses phonotactics outperforms one that utilizes only local distributional information. Moreover, the type of phonotactic information matters: a model that uses phonetically natural phonotactics outperforms one that uses all available phonotactic information.

In the first half of the paper (Section 2), I report a modeling study that was used to test which factors best predict reanalysis in Samoan. As a preview, Samoan reanalysis generally obeys local distributional patterns, but is also sensitive to a phonotactic restriction against sequences of homorganic consonants. These findings serve as the basis for an acoustic study (Section 3), where the phonotactic restriction against sequences of homorganic consonants is found to be phonetically motivated.

1.1. Local distributional information in the learning of paradigms

When speakers are faced with variable patterns in a paradigm, they are known to apply these patterns in a way that matches the proportion at which they occur within that paradigm. This type of sensitivity to local distributional information is often called FREQUENCY-MATCHING. For example, in Dutch, word-final obstruents are voiceless, but may alternate in voicing under suffixation. Thie means that given a stem ending in [t], this final [t] may either alternate with [d] under suffixation, as in (1a), or not alternate, as in (1b).

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

(1) *Dutch voicing alternations*

	STEM	SUFFIXED		
a.	[vɛrˈʊɛi**t**]	[vɛrˈʊɛi**d**-en]	'widen'	(alternating)
b.	[vɛrˈʊɛi**t**]	[vɛrˈʊɛi**t**-en]	'reproach'	(non-alternating)

Ernestus and Baayen (2003) find that when speakers are given nonce words and asked to produce their suffixed forms, they do so in a way that frequency-matches the rates of voicing alternation within the paradigms. For example, in the Dutch lexicon, final [f] alternates with [v] around 70% of the time. Speakers match this pattern, and apply voicing alternations to most [f]-final nonce words.

Frequency-matching (i.e. matching of statistical patterns local to a paradigm) 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 (2002); 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).

Moreover, existing models of reanalysis (and more generally, models of morphophonological learning) tend to be based on frequency-matching, relying only on local distributional information. 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).

1.2. The role of phonotactics in learning alternations

It is reasonable to posit that phonotactics can influence reanalysis, for the following reasons. Crosslinguistically, there tends to be a strong connection between phonotactics and paradigm-internal phonological patterns. In other words, alternations are usually consistent with stem phonotactics (Chomsky and Halle, 1968; Kenstowicz, 1996). This is especially true once we consider gradient phonotactics; Chong (2019) shows that even in cases of apparent mismatch between phonotactics and paradigm-internal alternation patterns, there is often some gradient phonotactic support for the alternation pattern. Additionally, alternations that are not supported by phonotactics tend to be under-attested.

Relatedly, many theories of acquisition 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 Artificial Grammar Learning experiment, Chong (2021) trains speakers on both a novel phonotactic pattern and novel alternation patterns. Results suggest that speakers draw on phonotactics to learn alternation patterns. 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).

Notably, it is unclear exactly how these two factors–phonotactics and paradigm-internal frequency distributions–interact. In particular, where there is a mismatch between phonotactics and local distributions, what do speakers do? This is a difficult question to address, as there are

relatively few languages with a clear case of mismatch between phonotactics and alternations. Experimental work that addresses the effect of phonotactics on alternations, such as Chong (2021) described above, have generally focused on exceptionless alternation patterns (where paradigm-internal distributions are unambiguous). The findings of the current study will therefore not just enrich our understanding of reanalysis, but also help us understand how phonotactics interacts with paradigm-internal frequency distributions.

1.3. Phonetic naturalness in phonological learning

Work on paradigm learning shows that in addition to frequency distributions, speakers are sensitive to various learning biases, and preferentially acquire patterns that are more 'natural'. This can result in the over-learning of more natural patterns (e.g. Kuo, 2023a), or under-learning of unnatural patterns (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).

My empirical focus will be on this second type of bias, a phonetic bias (or substantive bias). Effects of phonetic naturalness have been substantiated in various experimental work. For example, Wilson (2006) and White (2014) both find that when trained on novel alternation patterns, people preferentially learn ones that involve a phonetically smaller change. For example, White (2014) finds that the learnability of alternation patterns is directly correlated with the gradient similarity of sounds (measured using confusability experiments). Thus, a pattern where [b] alternates with [v] is easier to learn than one where [p] alternates with [v], because [b] is more phonetically similar to [v].

Based on these findings, it is important to consider how phonotactic information interacts with constraints on phonetic naturalness. As previewed above, I find that Samoan reanalysis is not just sensitive to phonotactics, but also constrained by phonetic naturalness.

2. Modeling reanalysis in Samoan

Samoan is an Oceanic language of the 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 words are always vowel-final. However, in suffixed forms, a consonant of unpredictable quality may surface, as seen by the examples in (2) using the ergative suffix. These alternations, also called **thematic consonant** alternations in the literature, are the focus of the current paper.

(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'

In this section, I present the results of a modeling study, where I find that reanalysis of Samoan thematic consonants is sensitive to a specific phonotactic constraint against sequences of homorganic consonants. Sections 2.1-2.3 describes the empirical patterns of reanalysis. Following this, Sections 2.4-2.6 will describe the modeling methodology and results.

2.1. Background

This section provides an overview of Samoan phonology, focusing on thematic consonant alternations, and the diachronic sound changes that make it possible for us to trace Samoan back to its pre-reanalysis state. Unless otherwise noted, descriptive generalizations are taken from Mosel and Hovdhaugen (1992). Additionally, results are based off of the *tautala lelei* register of speech, which preserves more segmental contrasts than the other register, *tautala leaga*. I focus on the *tautala lelei* register as it is the subject of most existing scholarly work.[2]

Samoan syllables follow a (C)V(V) structure; no codas or consonant clusters are allowed and onsets are optional. Samoan has five vowels /a, e, i, o, u/, all of which also show a two-way length contrast. 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. Additionally, /r/ is often realized as [l] even in careful speech.

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

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

Samoan thematic consonant alternations are observed in a variety of suffixal contexts, but I focus on the **ergative suffix**, as it is the most productive one. The ergative suffix has many allomorphs, split roughly into vowel-initial ones (/-a/, /-ina/, /-ia/), and ones which have a thematic consonant (/-Cia/ and /-na/, where 'C' is one of /f, m, t, s, l, ŋ, ʔ/. Examples of each allomorph are given in Table 2.

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, which Samoan belongs to. The relationship of Samoan to other Oceanic languages is summarized in Fig. 1, which shows a very simplified subgrouping of the Oceanic languages (Lynch et al., 2002; Pawley et al., 2007). Note that while there is some disagreement about the exact subgroupings, the general grouping of Polynesian languages under Oceanic is well-established.

For the languages affected by final consonant deletion, stem-final consonants were lost in unsuffixed forms but maintained in suffixed forms, resulting in unpredictable thematic consonant alternations (e.g. Proto-Oceanic **inum/*inum-ia* → Samoan *inu/inu-mia* 'to drink' and Proto-Oceanic **suat/suat-ia→sua/sua-tia*).

[2]The main difference between the registers is that *tautala lelei* preserves certain loanword phonemes, which I later exclude from my analysis. Additionally, Mosel and Hovdhaugen (1992) report that speakers are able to fluently switch between the two registers.

Erg.	stem	suffixed	gloss
a	lele	lele-a	to fly
ia	nofo	nofo-ia	to live, dwell
ina	iloa	iloa-ina	to see, perceive
sia	laka	laka-**s**ia	to step over
tia	pulu	pulu-**t**ia	to plug up
ŋia	tutu	tu-ŋia	to light a fire
fia	utu	utu-**f**ia	to draw water
mia	inu	inu-**m**ia	to drink
lia	tautau	tautau-**l**ia	to hang up
na	ʔai	ʔai-**n**a	to eat
ʔia	momo	momo-ʔia	to break in pieces

Table 2: Samoan thematic consonant alternations

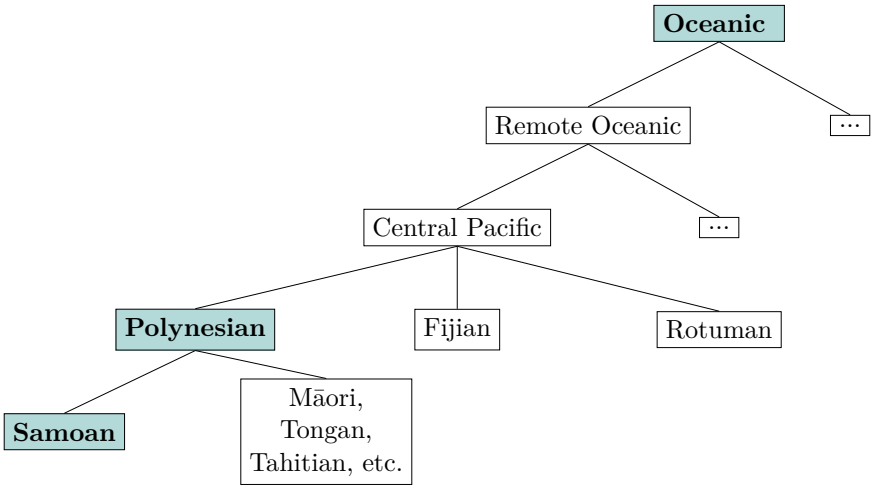

Figure 1: Internal subgrouping of Oceanic languages (Lynch et al., 2002; Pawley et al., 2007)

Crucially, not all Oceanic languages underwent final consonant loss. By comparing these languages, we can reconstruct what Samoan thematic consonant alternations would have looked like pre-reanalysis. Specifically, Proto-Oceanic (POc), the reconstructed ancestral language for Samoan, can be used as a proxy for what Samoan would have looked like pre-reanalysis; comparison of POc with Samoan can also give us insight into the patterns of reanalysis.

In general, if there has been no reanalysis, stems that historically ended in vowels (and in a subset of consonants) should take a vowel-initial suffix (/-a/, /-ia/, /-ina/). Otherwise, the suffix that surfaces is of the form /-Cia/, where /C/ should correspond to the historical POc stem-final consonant.[3] For example, [lele]∼[lele-a] 'to fly' descended from POc *rere, which is vowel-final. On the other hand, [inu]∼[inu-**m**ia] 'to drink' descended from POc *inu**m**, which ended in *m. Additionally, the relative distribution of the vowel-initial allomorphs (/-a/, /-ia/,

[3]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. The /-ina/ here is homophonous with the vowel-initial /-ina/ allomorph.

/-ina/) is prosodically conditioned; a discussion of these patterns is beyond the scope of the current paper, but can be found in Kuo (2023b).

Where there is a mismatch between POc and Samoan, this suggests that reanalysis has occurred. Some examples of this type of reanalysis are given in Table 3. For example, [aʔo] 'learn, teach' descends from POc *ako**t**, so its suffixed form should be [aʔo-**t**ia], but instead [aʔo-ina] is observed. This suggests that the allomorph has been reanalyzed from /-tia/ to /-ina/ (i.e. in the direction of t→∅).

		suffixed			
POc	stem	expected	actual	Reanalysis	gloss
*qato**p**	qato	ato-**f**ia	ato-a	f→∅	'thatch'
*ako**t**	aʔo	aʔo-**t**ia	aʔo-ina	t→∅	'learn, teach'
*puri	fuli	fuli-a	fuli-**s**ia	∅→s	'turn (over)'

Table 3: Examples of thematic consonant reanalysis in Samoan

2.2. Data: trends in the reanalysis of Samoan thematic consonants

In this section, I summarize the patterns of reanalysis in Samoan, using comparison of POc and Samoan forms. For simplicity, I combine the vowel-initial allomorphs, ignoring the factors that influence the relative distribution of /-a/, /-ina/, and /-ia/.

POc protoforms (n=1023) are taken from the Austronesian Comparative Dictionary (ACD; Blust et al., 2020). Items were excluded if they had fewer than six cognates within Oceanic. Modern Samoan forms are taken from the Milner (1966) dictionary and supplemented with forms from Pratt (1862/1893). I focus on stem-ergative pairs, since the ergative suffix is the most productive of all the suffixes that result in thematic consonant alternations. The resulting wordlist has 593 stem-suffix pairs.

The first trend we can observe is that reanalysis appears to generally be towards the locally most frequent allomorph; this is in line with the predictions of frequency-matching models of reanalysis. Fig. 2 compares the distribution of allomorphs in POc and Samoan, where POc represents pre-reanalysis Samoan. Historically, the majority of stems took vowel-initial allomorphs (n=704/1023, 69%). In modern Samoan, the proportion of vowel-initial allomorphs is roughly the same, but has increased slightly (n=425/593, 72%). Note that the modern Samoan data may under-estimate the proportion of stems which take /-a/ and /-ina/, since loanwords and other innovative forms that are omitted from the data will generally take /-ina/ (Mosel and Hovdhaugen, 1992). Nevertheless, the results suggest that reanalysis largely maintained the distribution of allomorphs present historically, and was otherwise towards the more frequent vowel-initial variants.

However, the distribution of ergative allomorphs is also conditioned by the identity of the preceding consonant. Moreover, reanalysis fails to match some of these consonant-conditioned distributional patterns, showing behavior that is *not* frequency-matching.

Fig. 3 compares the distribution of ergative allomorphs in POc and Samoan by identity of the preceding segment. For example, a cell where the preceding consonant is [**p**] and the allomorph is /-t̲ia/ represents suffixed forms like [i**p**o-t̲ia]. For ease of reading, the vowel-initial allomorphs are omitted, and the POc data is grouped by what the modern Samoan consonant would be (i.e. reflects the regular sound changes that occured between POc and Samoan).

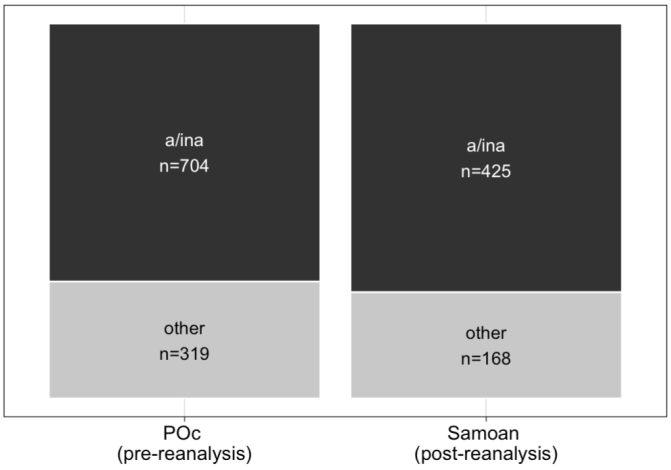

Figure 2: Distribution of ergative allomorphs before and after reanalysis

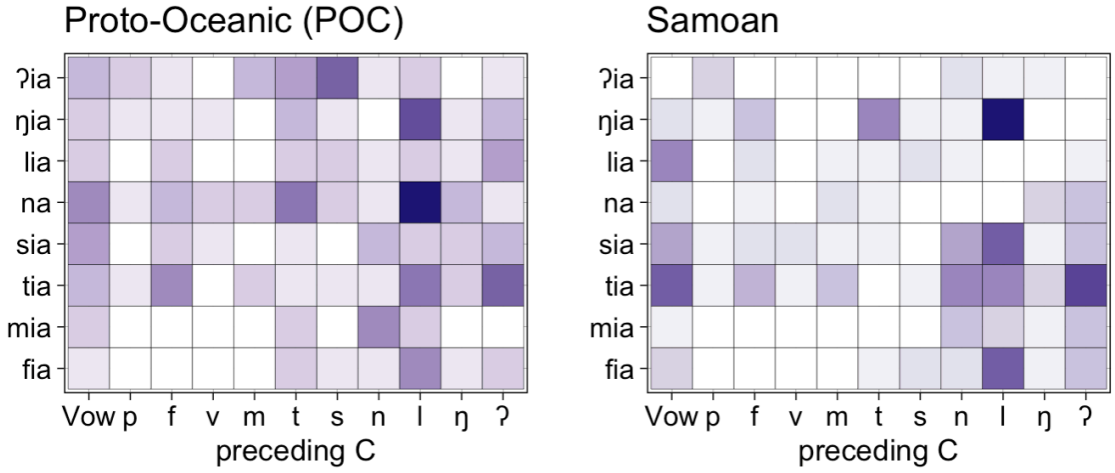

Figure 3: Distribution of ergative suffix allomorphs by preceding consonant in POc vs. Samoan

Some regularities in POc are maintained in Samoan. For example, in both POc and Samoan, when the preceding consonant is a labial (/p, f, v, m/), the ergative allomorph never starts with a labial (/-fia/, /-mia/). In other cases, however, there is a mismatch between the POc and Samoan distributions. In particular, stems of the type [i**l**o-n̲a] (where the suffix allomorph is [n̲a], and the preceding consonant is [**l**]) are relatively frequent in POc (n=11), but never attested in Samoan. In a Monte Carlo test of significance (detailed in Kuo 2023b, p. 117), I find that suffixed forms with sequences of coronal sonorants (e.g. [i**l**o-n̲a], [i**n**o-l̲ia]) are underrepresented in modern Samoan, given their historical distribution.

I argue that this mismatch is a result of reanalyses that are motivated by a phonotactic restriction in Samoan. Specifically, as will be discussed below in Section 2.3, Samoan has a dispreference for sequences of homorganic consonants (separated by an intervening vowel), and suffixed forms which have the violating sequences are more likely to be reanalyzed. Thus, suffixed forms like [i**l**o-n̲a] were disproportionately targeted for reanalysis because [l] and [n] are both coronals.

2.3. Data: Samoan stem phonotactics

A phonotactic dispreference for combinations of homorganic consonants in proximity to each other is accounted for in OT using OCP-place (Obligatory Contour Principle for Place of Articulation) constraints (McCarthy, 1988, 1994). Section 3.1 discusses functional motivations for OCP-place (and more generally the avoidance of sequences of similar segments). In this section, I present evidence that both Samoan and its historical predecessor have OCP-place restrictions.

In a detailed and comprehensive study of Samoan phonotactics, Alderete and Bradshaw (2013) find gradient OCP-place effects between consonants separated an intervening vowel. 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, where OCP-place effects are stronger for coronals which share the same manner of articulation. For example, [**n**u**l**a] is worse than [**t**u**l**a], because [n] and [l] are both sonorants, while [t] is an obstruent. In Kuo (2023b), I replicated Alderete and Bradshaw's results, with some methodological modifications. Findings are summarized here and discussed in more depth in Kuo (2023b).

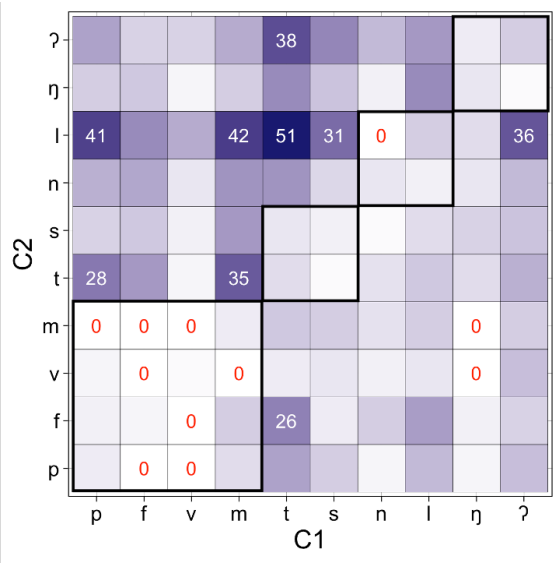

Figure 4: Consonant-consonant co-occurrences in Samoan

Figure 4 shows counts of all transvocalic consonant-consonant sequences (i.e. C_1VC_2) in Samoan. The data is taken from Alderete and Bradshaw (2013), who compiled monomorphemic headwords (i.e. unbounded roots) from the Milner (1966) dictionary. A total of 1,498 roots were analyzed (after excluding loanwords, classificatory names, and pseudo-reduplicated forms). 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). The outlined diagonal marks regions where C1-C2 co-occurrences tend to be less frequent. In particular, there appears to be a strong dispreference for labial-labial sequences and a dispreference for coronal-coronal sequences which share the same sonorancy (e.g. [s...t], [n...l]). Crucially, these are all regions that violate the OCP-place restriction. Additionally, [ŋ] and [ʔ] appear to pattern together, and sequences of [ŋ] and [ʔ] are also relatively infrequent. This pattern is not

well-motivated in modern Samoan (and is in this sense an acceintal gap). However, [ʔ] was historically the velar stop [k], meaning that at some point, [ŋ] and [ʔ] were homorganic (and both velar).

There are some other C1-C2 sequences that are underrepresented. For example, [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).

Following Wilson and Obdeyn (2009), the effect of OCP-place was confirmed in a probabilistic phonotactic model, where different phonotactic restrictions are encoded as constraints. This method allows for statistical testing of OCP-place effects after controlling for the baseline frequency of each consonant. Table 4 lists the phonotactic constraints that were tested; in addition to a general OCP-place constraint (which assigns violations to any two homorganic C1-C2 pairs), I tested place-specific constraints; for example, OCP-LABIAL assigns violations to homorganic C1-C2 pairs only if they are labial. Finally, because OCP-place effects are often stronger when the target segments also share other similarities (e.g. Frisch, 1996; Coetzee and Pater, 2006; Wilson and Obdeyn, 2009), constraints that additionally care about sonorancy are included (e.g. OCP-LABIAL-SON assigns violations to homorganic C1-C2 pairs only if they are labial *and* share the same sonorancy).

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). Results match the qualitative observations discussed above: OCP-place effects are present across all places of articulation; for the coronals, they are additionally conditioned by sonorancy.

CONSTRAINT	EX. VIOLATIONS	ΔL	P
OCP-place	pama, tala, nala	-0.01	n.s.
OCP-LAB	pama, pava, papa	**6.03**	**0.0005*****
OCP-LAB-SON	mama, papa, pafa	1.54	n.s. (0.08)
OCP-COR	tasa, tasa, tala	-0.01	n.s
OCP-COR-SON	nala, lala, tasa	**34.76**	**7.56×10^{-17}*****
OCP-BACK	ŋaʔa, ʔaʔa,ŋaŋa	**3.94**	**0.002****
OCP-BACK-SON	ŋaŋa, ʔaʔa	0.01	n.s.

Table 4: OCP constraint weights learned by the phonotactic model

Notably, while OCP-place effects are present in modern Samoan phonotactics, there are also strong reasons to believe that they were present in an earlier stage of Samoan, and therefore were able to influence reanalysis. First, OCP-place effects have been documented across multiple Polynesian languages, leading Krupa (1966, 1967, 1971) to posit that they were present in Proto-Polynesian (the reconstructed language from which Polynesian languages, including Samoan, descend from).

In fact, in a corpus of Proto-Polynesian (PPn) protoforms, I find the same OCP-place effects that are present in modern Samoan. Fig. 5 shows consonant-consonant co-occurrences in Proto-Polynesian. Counts are based off of a corpus of 1645 protoforms taken from the the Polynesian Lexicon Project (POLLEX-Online; Greenhill and Clark, 2011). For comparability with the Samoan

data, consonants are grouped by their corresponding sound in modern Samoan.

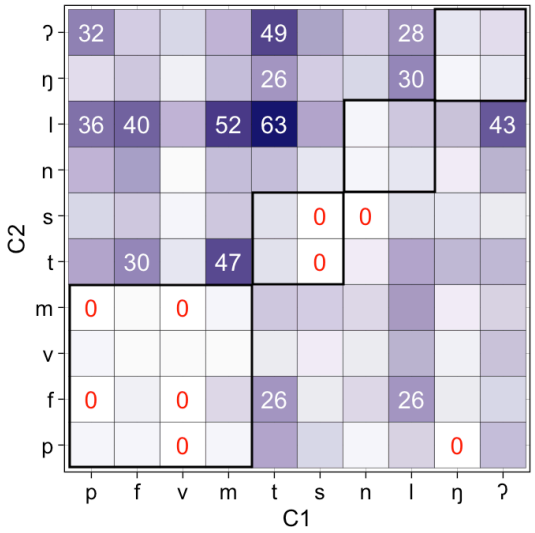

Figure 5: Consonant-consonant co-occurrences in PPn

The boxes frame regions where OCP-place effects were found for Samoan, and therefore where C1-C2 pairs are expected to be underrepresented. We can see that in general, the PPn distributions match the Samoan distribution. This was confirmed using the same phonotactic model described above. The results, given in Table 5, are consistent with the findings for the modern Samoan data. In particular, OCP-LAB, OCP-COR-SON, and OCP-BACK tested as significant, and these were the same three constraints found to be significant for Samoan.

CONSTRAINT	ΔL	P
OCP-place	0.005	n.s. (0.92)
OCP-LAB	**33.83**	**1.95×10^{-16}*****
OCP-LAB-SON	0.03	n.s. (0.81)
OCP-COR	0.99	n.s (0.16)
OCP-COR-SON	**30.15**	**8.12×10^{-15} * ****
OCP-BACK	**3.97**	**0.005****
OCP-BACK-SON	0.01	n.s. (0.36)

Table 5: OCP constraint weights learned by the phonotactic model for Proto-Polynesian

2.4. Methodology: model architecture

Although existing models of reanalysis are frequency-matching (i.e. match local patterns), Samoan reanalysis appears to also be sensitive to an OCP-place phonotactic restriction. In particular, suffixed forms that violate OCP-place are absent more than predicted by frequency-matching. To test this hypothesis, I implement a quantitative model of reanalysis.

To formally implement the interaction between frequency-matching and phonotactics, I use Maximum Entropy Harmonic Grammar (MaxEnt; Goldwater and Johnson, 2003; Wilson, 2006), which is a probabilistic model of phonological learning that uses weighted constraints. A preference for outputs that obey the phonotactics can then be implemented as a prior (see also Wilson 2006 and White 2013, 2017 for similar implementations).[4]

In principle, other probabilistic phonological models could be used. However, MaxEnt is well-suited to the type of learning behavior being modeled, because its general mechanism of weighting constraints according to the principle of maximum entropy results in frequency-matching. More concretely, the model is trained on stem-ergative paradigms, and will match the frequencies of this data. The subsequent addition of a prior allows for us to model frequency-matching that is constrained by phonotactics.

The model inputs were 500 stem-ergative paradigms whose distribution reflect the historic (POc) frequencies. The model inputs reflect several simplifying assumptions. First, the vowel-initial allomorphs (/-a/,/-ia/,/-ina/) 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'). For example, a stem-ergative pair like [i**n**o]∼[i**n**o-**l**ia] reflects all forms where the preceding consonant is [l] and the suffix allomorph is /-lia/.[5]

Recall that phonotactics are the global statistical regularities of a language, and a phonotactic bias is essentially a tendency to obey these regularities. To implement this bias, I first train phonotactic grammars on monomorphemic Samoan roots. The phonotactic model I adopt is the UCLA Phonotactic Learner (UCLAPL; Hayes and Wilson, 2008), which is itself a MaxEnt grammar that learns phonotactic constraint weights in a way that matches the probabilities of the lexicon. The input to the phonotactic model is a corpus of 1645 Proto-Polynesian protoforms taken from POLLEX (Greenhill and Clark, 2011). The corpus was modified to reflect the regular sound changes that have happened between Proto-Polynesian and Samoan, and is meant to reflect the phonotactics of an earlier stage of Samoan, pre-reanalysis.

The UCLAPL, once trained, can assign penalty scores to new words (where a higher penalty means that a word is phonotactically worse). I use the phonotactic grammar to assign penalty scores to the suffixed forms of the model of reanalysis (e.g. [ilo-tia], [ino-lia], [ipo-fia], etc.). These penalty scores then become the basis for a constraint UsePhonotactics, that is put into the model of reanalysis and given a bias towards higher weight.

In implementing a phonotactic bias, we can also consider which statistical regularities speakers utilize for reanalysis. Speakers might simply be sensitive to segment-segment combinations in the language, or they might generalize to phonologically active classes. Here, I follow Mielke (2008)

[4]Details of the model architecture are not the focus of the current paper, but the reader is referred to Kuo (2023b) for a more thorough discussion.

[5]Note that as famously pointed out by Hale (1968, 1973), the Polynesian thematic consonant can be analyzed as underlyingly belonging to the stem, or to the suffix allomorph, as I have assumed in this paper. Note however that reanalysis can be modeled in both approaches; Kuo (2023b, Ch. 4.4.1) discusses the motivations for the choosing the current approach, as well as ways to analyze reanalysis in the other approach.

and adopt to term "phonotactically active class" to refer to groups of sounds that pattern together phonologically, but are not necessarily natural classes in the sense of being phonetically motivated.

Additionally, speakers might pick up on any statistical regularities in the language, or they may be constrained to only pick up on 'principled' generalizations that are motivated by factors like phonetic naturalness. To test between these different possibilities, I implement three phonotactic grammars; each grammar has constraints on C1-C2 combinations (separated by an intervening vowel), but they have different assumptions about what phonotactic constraints should look like:

1. BIGRAM MODEL: This model was trained on a constraint set that consisted of all possible C1-C2 combinations, where C1 and C2 are segments. For example, the constraint *p...l penalizes forms like [**p**a**l**a] and [i**p**o**l**ia].
2. ACTIVE CLASS MODEL: This model learned 50 constraints inductively, and but was not otherwise given pre-specified constraints. As a result, it can learn constraints on phonologically active classes.
3. OCP MODEL: This model was trained on a set constraints set that included all possible combinations of OCP-place (OCP-LABIAL, OCP-CORONAL, and OCP-BACK), crossed with the features [sonorant], [voice], and [continuant]. For example, constraints on labial-labial sequences include OCP-LABIAL, OCP-LABIAL-son, OCP-LABIAL-voice, and OCP-LABIAL-continuant. Although this model generalizes to natural classes, it is more restrictive than the NATURAL CLASS MODEL in that it can only learn constraints on homorganic consonants.

The ACTIVE CLASS and OCP models both allow generalization to phonologically active classes, while the BIGRAM model doesn't. If the BIGRAM model performs as well as or outperforms the other models, this suggests that speakers are simply learning C1-C2 probabilities and applying this information to resolve ambiguities in an alternation pattern. On the other hand, if the ACTIVE CLASS and OCP models perform better, this suggests that speakers prefer to generalize patterns to classes of sounds. The OCP model is additionally more restrictive, in that it only allows for phonetically motivated OCP constraints, rather than potentially arbitrary constraints learned over any group of segments. If the OCP model outperforms the other models, this suggests that speakers do not utilize all phonotactic regularities in the lexicon, but prefer to learn more well-motivated constraints.

The model as described so far is able to match frequencies of the input data, but in a way that is constrained by phonotactics. It should additionally be able to simulate the effect of reanalyses over time. To do this, I adopt an iterated learning paradigm, where one iteration of the model becomes the input to the next iteration. Under this approach, small changes to an alternation pattern can accumulate over iterations (each taken to be a generation of speakers), resulting in large-scale reanalyses of a pattern. I assume a relatively simple agent-based architecture, where each generation has just one speaker and one learner, but other approaches include: 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 variants of Optimality Theory (e.g. Boersma, 1998; Zuraw, 2003).

The iterated learning approach I use is illustrated in Fig. 6. In the first iteration, the speaker S1 produces the output language based on their grammatical knowledge (i.e. Grammar 1; G_1). The grammar is a MaxEnt phonological grammar. The learner observes these data, induces the relevant generalizations, and forms another grammar (G_2), which then becomes the basis of the output data presented to the next generation. This process is repeated for many iterations.

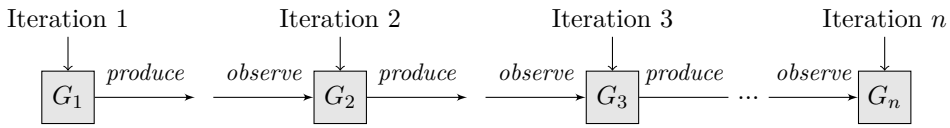

Figure 6: Structure of an iterated learning model, adapted from Ito and Feldman (2022, p. 3). H_i indicates hypotheses of each generation.

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 learner "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 learner's grammar.

Note that this simplified approach does not consider the interaction of multiple speakers, 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 component 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.

2.5. Results

Three phonotactically-biased models were compared; in these models, the UsePhonotactics constraint was biased to have higher weight than other constraints. The three models differ in the phonotactic model that was used (Bigram, Active Class, OCP), but are otherwise identical. Each model is also compared against a corresponding Baseline model, which has the same constraints but no phonotactic bias; specifically, the prior prefers all constraints to have the same weight.

A good model of reanalysis should, when given the historical Samoan pattern, be able to predict reanalysis towards the modern Samoan pattern. As such, models were evaluated on how well they fit the *modern* Samoan data. Table 6 compares the log-likelihood of each model, fit to modern Samoan. The baseline models are combined because they had nearly identical performance. A higher (less negative) log-likelihood indicates better model fit. The rightmost column (ΔL) shows the change in log-likelihood of each model compared to the baseline.

Overall, all four phonotactically-biased models outperform the baseline model. Of the these three models, the bigram model performs the worst, while the OCP grammar has the best performance. The Active class and OCP grammars both generalize to classes of sounds, but the OCP grammar does better.

A closer inspection of the data shows that this is because the phonotactically-biased models perform better than the baseline in predicting reanalysis for forms like [ino]∼[ino-lia], which

	L	ΔL
BASELINE	-2448.81	–
ACTIVE CLASS	-2416.27	32.54
OCP	**-2385.00**	**63.81**
BIGRAM	-2438.39	10.42

Table 6: Model results: log likelihood

involve an OCP-place violation. For example, Fig. 7 compares predictions of the BASELINE and OCP model for stems with a preceding [l] (i.e. of the type [ilo]). For ease of interpretation, only a subset of stem-ergative pairs 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.

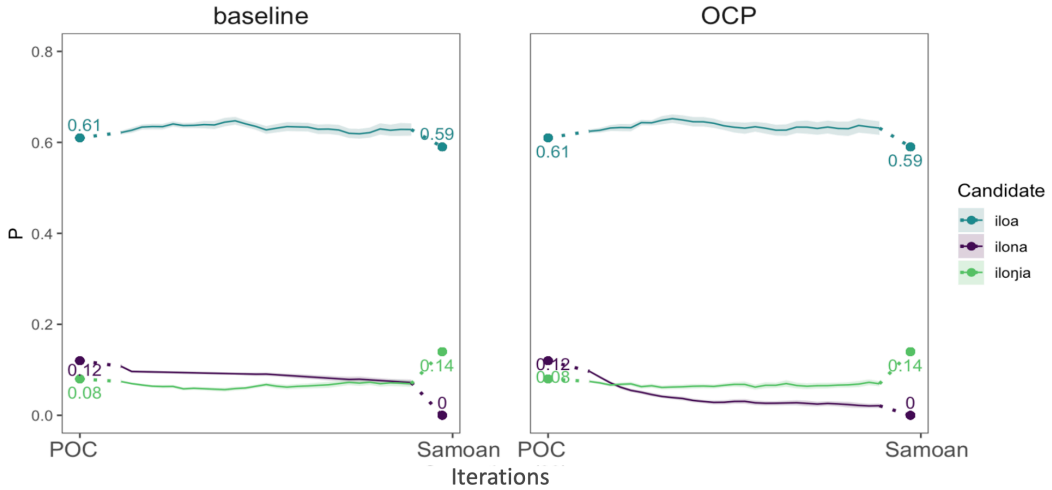

Figure 7: Model predictions in stems with a preceding /l/

2.6. Discussion

As shown above, all phonotactically-biased models outperformed the BASELINE models. This supports the hypothesis that reanalysis of Samoan thematic consonants is generally frequency-matching, but constrained by phonotactics.

The relative performance of the three phonotactically-biased models also gives us insight into what generalizations speakers pick up on. The BIGRAM model performs the worst, suggesting that models which generalize to phonologically active classes are better predictors of learner behavior. Additionally, the OCP model outperforms the ACTIVE CLASS model. This is likely because the ACTIVE CLASS grammar learns constraints that are not sufficiently general, especially for the coronal sonorants.

In the model inputs (i.e. historical, pre-reanalysis Samoan), words of the type [i**n**o-**n**a], [i**n**o-**l**ia], and [i**l**o-**l**ia] were infrequent, but [i**l**o-**n**a] stems were relatively frequent. The NATURAL CLASS grammar, which inductively finds constraints from data, therefore learned the three specific constraints given in (4), rather than a general OCP-COR-SON constraint. The constraint *[l...n] is assigned a relatively lower weight, so the model does not penalize [ilo-na] type words as heavily

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

(4) *Constraints on coronal sonorant C1-C2 pairs in the* ACTIVE 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

Overall, comparison of the different phonotactic grammars suggests that speakers do not utilize all phonotactic regularities in the lexicon. Instead, speakers are picking up on OCP-place constraints. Moreover, OCP-place effects in Samoan are gradient, where for the coronal sounds, OCP-place is much stronger when the target consonants match in sonorancy. In the next Section, I will argue that both facts fall out from the phonetic motivation behind OCP-place.

3. The phonetic naturalness of OCP-place

In Section 2, findings from a model of reanalysis suggest that in Samoan, reanalysis is constrained by OCP-place. Notably, while OCP-place had an effect on reanalysis, other phonotactic regularities did not. In this section, I propose that this is because reanalysis is further constrained by phonetic naturalness. In particular, I argue, following work by Frisch (1996); Frisch and Zawaydeh (2001), that OCP-place is rooted in phonetic similarity avoidance, and present the results of an acoustic study which supports this analysis.

Note that the acoustic study focuses on the labials and coronals, and does not consider /ŋ/ and /ʔ/. This is because /ŋ/ and /ʔ/ form a class of size two, so meaningful comparisons of gradient similarity are not possible. Additionally, /ʔ/ is often elided in natural speech (Mosel and Hovdhaugen, 1992), and is difficult to segment due to its highly variable realization.

Section 3.1 will give an overview of the literature on OCP-place, with a focus on arguments that OCP-place is phonetically motivated. Following this, Sections 3.2-3.5 present the results of an acoustic study that quantifies consonant similarity in Samoan using measures of spectral similarity.

3.1. Background

OCP-place effects are well attested crosslinguistically. 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 often, 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 or more of a set of non-place features (McCarthy, 1988; Yip, 1989; Padgett, 1991, 1995; Wilson and Obdeyn, 2009). In Arabic, like for Samoan, OCP-place effects are stronger for coronals that share the same sonorancy (Pierrehumbert, 1993; Frisch and Zawaydeh, 2001). More concretely, sequences like [t...d] and [n...l] are more marked than [t...l] and [n...d]. As pointed out by Pierrehumbert (1993), this gradience makes OCP-place effects difficult to account for in non-probabilistic grammars.

Frisch (1996) and Frisch et al. (2004) argue that the gradience of OCP-place is a direct consequence of OCP-place being rooted in a functional phonetic motivation. Specifically, people 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 replicated across many psycholinguistics studies. For example, the repetition of like segments in close proximity has been known to increase speech error rates (Dell, 1984) and overall production rate (Sevald and Dell, 1994). Similar types of processing difficulties have been reported in perception tasks (e.g. Miller and MacKay, 1994). In work on Arabic, Berg and Abd-El-Jawad (1996) find that words with OCP-place violations are more susceptible to speech errors involving consonant misordering.

Consistent with these studies, Frisch (1996) and Frisch et al. (2004) find that the strength of OCP-place in Arabic directly correlates with measures of consonant similarity. In these studies, they use phonological features 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).

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

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 (Mielke, 2008). Feature-based measures also ignore the variable phonetic realization of target phones. In Samoan, for example, [v] is variably lenited and may therefore be closer to sonorants in its phonetic realization.

As an alternative, I propose that phonetic similarity can be quantified as the spectral distance between two phones. This method is described in Section 3.2.

3.2. Methods

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.

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; Mielke, 2012; Cychosz et al., 2019; Cychosz, 2022).

MFCCs are well-suited to the current task of measuring consonant-consonant similarity because they measure the overall shape of the spectrum, allowing for comparability between a broader range of consonant manners. In contrast, traditional measures such as formant tracking are more sensitive to tracking errors and often not comparable across different manners of consonants.

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

Spectral distance between two consonants C1 and C2 was measured as the Euclidean distance between their average MFCCs using the equation in (6), where $\bar{x}_{C1}$ and $\bar{x}_{C2}$ are the averaged MFCCs of each segment. 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/.

(6) $d(C1, C2) = \sqrt{(\bar{x}_{C1} - \bar{x}_{C2})^2}$

3.3. Data

Data comes from audio recordings of three male speakers from the Jehovah's Witnesses website.[6] 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. However, compared to lab-collected speech, the dataset is also more naturalistic, and includes tokens 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

3.4. Results

In the Samoan phonotactic results (Section 2.3), the strength OCP-place was found to be conditioned on Place, such that OCP-place effects were stronger for labials than for coronals. In addition, there was an effect of sonorancy conditioned on place; for the coronals, OCP-place was stronger when the target segments shared the same sonorancy; this effect was not present for labials. Here, I test whether these same factors are good predictors of phonetic similarity between

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

consonants. Results are summarized in Fig. 8 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.

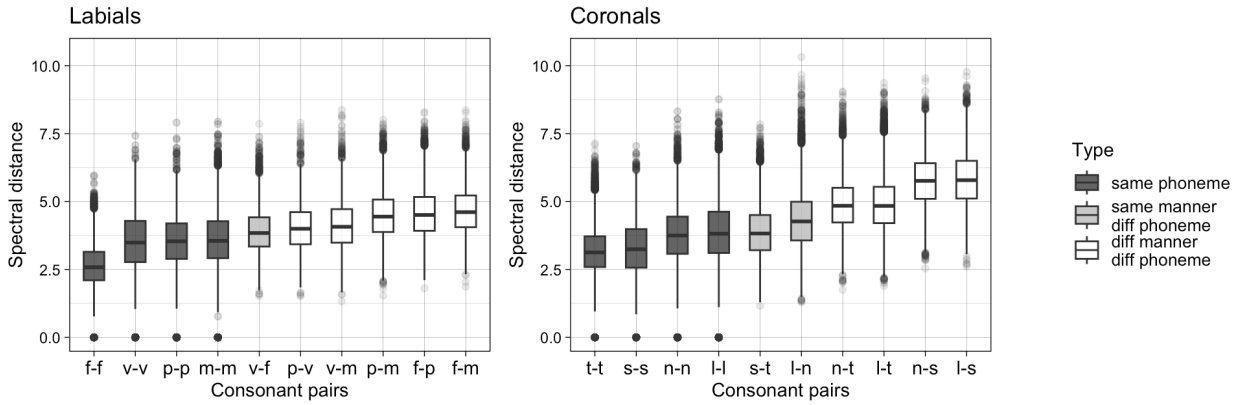

Figure 8: Spectral distances between consonant-consonant pairs

Looking at the figure, the spectral distances between labials are lower overall (compared to the distances between coronals), suggesting that they are acoustically more similar to each other. At the same time, spectral distances within the labials are more compact; there is less variation across different consonant-consonant pairs. The effect of manner is also weaker.

A linear mixed effects regression (LMER) model was used to confirm these observations. The model was conducted in R using the lme4 package (Bates et al., 2015), with speaker as a random effect and spectral distance as a dependent variable. Main effects of PLACE (labial vs. coronal), MATCH-SON (whether C1 and C2 match in sonorance, yes vs. no), and MATCH-SEG (whether C1 and C2 are identical, yes vs. no) were included. PLACE and MATCH-SON are included based on the phonotactics, as both were found to influence the strength of OCP-place effects. MATCH-SEG was also included to see if there was an additional effect of identity on consonant similarity. Additionally, I tested for the interaction of PLACE and MATCH-SON, and the interaction of PLACE and MATCH-SEG. If the strength of OCP-place effects is based on phonetic similarity, we should observe an interaction between PLACE and MATCH-SON, such that MATCH-SON has a stronger effect when the consonants are coronal.

Model results are summarized in Table 8. All effects were found to be significant in Likelihood Ratio Tests (performed using the *anova()* function). As expected, increase in segment similarity (indicated by MATCH-SON and MATCH-SEG) decreases spectral distance. Consistent with Fig. 8, there is an overall increase in spectral distance when the C1-C2 pair is coronal (β=0.71, CI = [0.70,0.73]). Finally, there is also a significant interaction of PLACE and MATCH-SON, such that matching sonorancy results in a greater decrease in spectral distance for coronals (vs. labials).

3.5. Discussion

Results suggest that spectral similarity corresponds closely to the OCP-place trends in the lexicon. In the lexicon, there is a strong effect of OCP-LAB, but not OCP-LAB-SON. The opposite is true for coronals, where there is an active OCP-COR-SON constraint but not a general OCP-COR constraint.

dist ~ PLACE+MATCH-SON+MATCH-SEG+PLACE:MATCH-SON+PLACE:MATCH-SEG + (1|speaker)

Predictors	*Estimates*	*CI*	*p*
Main effects			
(Intercept)	4.50	[4.42, 4.58]	<0.001
PLACE [cor]	0.71	[0.70, 0.73]	<0.001
MATCH-SON [yes]	-0.23	[-0.24, -0.21]	<0.001
MATCH-SEGMENT [yes]	-1.02	[-1.03, -1.00]	<0.001
Interaction effects			
PLACE [cor] x MATCH-SON [yes]	-0.86	[-0.88,-0.84]	<0.001
PLACE [cor] x MATCH-SEGMENT [yes]	0.25	[0.43,0.47]	<0.001

Table 8: LMER model results for predictors of spectral distance between two segments

This matches the spectral similarity data, where labials are overall more similar to each other (i.e. smaller spectral distance), in a way that is less sensitive to sonorancy. In contrast, coronals are overall less similar to each other, and there is a greater effect of sonorancy; coronal-coronal pairs that mismatch in sonorancy are acoustically more different (i.e. have a higher spectral distance) than ones that match in sonorancy.

Overall, the results of this acoustic study suggest that the gradience of OCP-place in Samoan is rooted in 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.

4. Conclusion

In this paper, I argued that reanalysis of stem-ergative paradigms in Samoan is constrained by both phonotactics and phonetic naturalness. This contrasts with previous models of reanalysis, which are primarily frequency-matching, meaning that they utilize only distributional information local to the paradigm. More generally, these results provide new evidence that fine-grained phonetic detail, in this case the phonetic similarity of consonants, can influence the learning of morphophonological paradigms.

In Section 2, a quantitative model of reanalysis was used to test between theories of how Samoan reanalysis occurred. I found that a model which incorporates phonotactics outperforms one that is purely frequency-matching. Importantly, the choice of phonotactics also appears to be constrained; a model restricted to learning just OCP-place effects outperformed ones that were able to learn any phonotactic regularities. I suggest that this is because reanalysis is further constrained by phonetic naturalness, and that OCP-place is rooted general processing constraints against the repetition of phonetically similar segments. Building on this proposal, Section 3 outlines an acoustic study where consonant similarity in Samoan was quantified using spectral distance measures. The results are consistent with the phonetic naturalness account, and patterns of consonant-consonant similarity are closely matched with the strength of OCP-place effects found in the lexicon.

Notably, as Glewwe (2019) points out, deviations from frequency-matching are hard to find in experiments. Where experimental work has found non-frequency-matching behavior, it has almost always been a preference for non-alternation. For example, people prefer paradigms like

[ra**t**]∼[ra**t**-e] over ones like [ra**t**]∼[ra**d**-e], because the latter paradigm involves a [t]∼[d] alternation. In contrast, the OCP-place effects found in the current study cannot be characterized as a preference for non-alternation. Experimental results on these type of effects may have been mixed because they are of such a small magnitude that they cannot be reliably found in an experimental setting. In these cases, data from language change can prove especially helpful; the ecological validity of this data makes it a suitable 'natural testing ground' for theories of linguistic learning.

More generally, 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) analysis of OCP-place in Arabic, where constraints are selected from a universal inventory of possible constraints, rather than a result of phonetically motivated diachronic changes.

References

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

Albright, A.C., 2002. 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.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67. doi:10.18637/jss.v067.i01.

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

Berg, T., Abd-El-Jawad, H., 1996. The unfolding of suprasegmental representations: a cross-linguistic perspective1. Journal of Linguistics 32, 291–324.

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.

Blust, R., Trussel, S., Smith, A.D., 2020. 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.

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.

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.

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.

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., 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.

Hyman, L.M., 1976. Phonologization, in: Juilland, A. (Ed.), Linguistic studies presented to Joseph H. Greenberg. volume 4, pp. 407–418.
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.
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.
Ling, C., Marinov, M., 1993. Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. Cognition 49, 235–290.
Lynch, J., Ross, M., Crowley, T., 2002. The oceanic languages. volume 1. Psychology Press.
MacWhinney, B., Leinbach, J., 1991. Implementations are not conceptualizations: Revising the verb learning model. Cognition 40, 121–157.
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., 2008. The Emergence of Distinctive Features. Oxford Studies in Typology and Linguistic Theory, OUP Oxford.
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.
Miller, M.D., MacKay, D.G., 1994. Repetition Deafness: Repeated Words in Computer-Compressed Speech Are Difficult to Encode and Recall. Psychological Science 5, 47–51. doi:10.1111/j.1467-9280.1994.tb00613.x.
Milner, G.B., 1966. Samoan Dictionary; Samoan-English, English-Samoan. ERIC.
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., Bedford, S., Sand, C., Connaughton, S., 2007. The origins of early lapita culture: the testimony of historical linguistics. Oceanic Explorations , 17–49.
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.
Ramsammy, M., 2015. The life cycle of phonological processes: Accounting for dialectal microtypologies. Language and Linguistics Compass 9, 33–54.
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.
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.
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.
Weinreich, U., Labov, W., Herzog, M., 1968. Empirical foundations for a theory of language change. University of Texas Press.
White, J., 2014. Evidence for a learning bias against saltatory phonological alternations. Cognition 130, 96–115.
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., 2003. Probability in language change, in: Bod, R., Hay, J., Jannedy, S. (Eds.), Probabilistic Linguistics. MIT Press, pp. 139–176.
Zuraw, K.R., 2000. Patterned exceptions in phonology. Ph.D. thesis. University of California, Los Angeles.