# 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 substance. 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 supported by an iterated learning model that is based in MaxEnt (Goldwater and Johnson, 2003). 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, and more generally that in reanalysis, speakers preferentially utilize phonetically-motivated phonotactics.

*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). The current paper investigates how phonetic detail can affect a specific type of change involving morphophonological paradigms. Paradigms in this context refers to sets of grammatically related forms which share a morpheme. For example, the English past-tense paradigm includes forms like *jump/jumped, heap/heaped, cook/cooked,* where each pair of words (e.g. *jump/jumped*) is paradigmatically related.

Paradigms can often have conflicting data patterns. Building on the example of English past tense formation, past tense can be formed in multiple ways (e.g. *want/wanted*, *bleed/bled*, *speak/spoke*, etc.). This is potentially challenging for learners (referring here to children tasked with learning of a language)[1], 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.

[1]Throughout the paper, unless referring to specific model implementations such as the UCLA Phonotactic Learner, I use 'learner' to refer to speakers acquiring a language, and 'model' to refer to computational models that aim to simulate aspects of this human learning

| **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).[2]

In this paper, I use data from Samoan verbal paradigms to address how reanalysis of morphophonological paradigms is constrained by i) phonotactics and ii) phonetic substance. Existing models of reanalysis tend to focus on how learners utilize probabilistic distributions within a paradigm. However, there is reason to believe that at least two other factors, phonotactics and phonetic substance, can also influence reanalysis. This paper therefore aims to address a gap in our understanding of how phonotactics and phonetics can shape morphophonology. These different factors–distributions within a paradigm, phonotactics, and phonetic substance–are discussed in Sections 1.1-1.3.

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. As such, findings from this paper not only improve our understanding of diachrony, but also have the potential to provide insight into how synchronic morphophonological learning works.

### *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. This 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) *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.

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

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).

There are two notable exception to frequency-matching that have been observed in the literature. First, Albright and Hayes (2003) find that patterns which are not well-attested (i.e. apply to very few forms) may be disfavored, and applied at a lower rate than would be expected in a purely frequency-matching model. Additionally, in the context of language acquisition, children tend to *over-regularize* patterns instead of frequency-matching (Marcus et al., 1992; Hudson Kam and Newport, 2005; Schumacher and Pierrehumbert, 2021). When this happens, learners typically generalize towards the most frequent pattern. In general, frequency-matching predicts change to be preservatory, while overregularization predicts reanalysis towards the more frequent variant. Importantly, even in these cases where speakers do not frequency-match, there is evidence that they track and utilize paradigm-internal frequencies.

### *1.2. The role of phonotactics in learning alternations*

Phonotactics is distributional knowledge about how segments can combine across the entire language, not just local to a paradigm. 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, morphophonological alternations are usually consistent with stem phonotactics (Chomsky and Halle, 1968; Kenstowicz, 1996). For example, in English plural formation, the plural alternates between [z], [s], and [ɪz] (e.g. dog/dog[z], cat/cat[s], fish/fish[ɪz]). This alternation is only observed with the plural suffix (and with a few homophonous suffixes), so in this sense is local to the noun paradigm. However, it is also motivated by a general phonotactic constraint which holds true across English, that obstruent-obstruent sequences must agree in voicing (e.g. 'duct' [dʌkt] is a legal English word, but *[dʌkd] is not).

The connection between phonotactics and alternations is even more prevalent 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).

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? Do speakers rely completely on paradigm-internal distributions, or do they also rely on phonotactics? These are difficult questions to address, as there are relatively few natural 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 alternation patterns that apply 100% of the time (meaning that the paradigm-internal distributions are never ambiguous).

In contrast, the present study focuses on a Samoan paradigm where no alternation applies exceptionlessly, and there is ambiguity in the paradigm-internal distributions. The findings of the current study will therefore not just enrich our understanding of reanalysis, but serve as a testing ground for understanding 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'. Here, naturalness can has a broad meaning, ranging from typologically common to phonetically motivated. Such biases 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: the first is complexity bias, or a bias against formally complex patterns, usually measured in terms of features (Moreton and Pater, 2012a). For example, a constraint *V[-sonorant]V would be less complex than *V[-sonorant, -labial]V, as the former uses fewer features.

The second type of bias is substantive bias, or a bias against phonetically unnatural patterns (Moreton and Pater, 2012b). Following Glewwe (2019), I adopt the view that a phonological pattern qualifies as phonetically natural if it can be shown to be phonetically motivated, where the motivation can be perceptual or articulatory. 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. In particular, White (2014) shows 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]. This bias is phonetically motivated, in the sense that it directly refers to the phonetic perceptibility of different alternation patterns.

My empirical focus will be on this second type of bias, a phonetic bias (or substantive bias). Existing experimental work supports the existence of phonetic bias in phonological learning, so it is important to consider how constraints on phonetic naturalness affect paradigm learning. I focus on phonetic bias (rather than complexity bias), because as will be discussed below, Samoan has a phonotactic restriction against sequences of homorganic consonants, analyzed in Optimality-Theoretic terms as OCP-place. OCP-place is often argued to be phonetically motivated; crosslin-

guistically, the strength of OCP-place is often gradient, in a way that appears to be directly related to the phonetic similarity of the consonants being targeted.

### *1.4. The present study*

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). In Samoan, many paradigms have ∅∼C alternations, which are also known as **thematic consonant** alternations in the literature; the table in (2) gives examples of these alternations using the ergative suffix.

(2) *Examples of Samoan thematic consonant alternations*

| STEM | ERGATIVE | GLOSS |
|---|---|---|
| pulu | pulu**t**ia | 'to plug up' |
| laka | laka**s**ia | 'to step over' |
| inu | inu**m**ia | 'to drink' |
| fulu | fulu-a | 'to rub, wash' |

Samoan ∅∼consonant alternations are mostly unpredictable. This means that when a Samoan speaker is given a stem they have never seen before, such as a hypothetical stem like *palu*, they won't know what the suffixed form should be (*palu-tia, palu-mia, palu-sia, palu-a,* etc.). This makes the task of learning Samoan paradigms challenging. As a result, it is likely that there has been extensive reanalysis of the alternation pattern over time.

Existing studies on reanalysis of Polynesian thematic consonant alternations have mostly focused on the effect of paradigm-internal frequencies. For example, Blevins (2008) finds that across various Polynesian languages, including Maori and Hawaiian, reanalysis tends to be towards the more frequent alternants. In this study, I explore the possibility that in addition to paradigm-internal frequencies, phonotactics and phonetic naturalness might also affect reanalysis.

First, in Section 2, I report a corpus and modeling study that tests whether phonotactic constraints are active in reanalysis. To preview the findings, I find that while phonotactics do appear to be active in reanalysis, not all phonotactic constraints are equally important. Instead, there are effects of a specific phonotactic restriction against sequences of homorganic consonants. This restriction, which I refer to as OCP-place, penalizes forms like [**n**a**l**u], which contains a sequence of two coronal consonants [n] and [l] (separated by a vowel).

One possible explanation for this finding is that while phonotactics can influence reanalysis, speakers preferentially pick up on the phonotactics that are phonetically motivated. This is plausible in the case of OCP-place, which has been argued to have a functional phonetic motivation, in that it is the avoidance of phonetically similar segments (Frisch, 1996; Frisch et al., 2004). However, the phonetic basis of OCP-place has not been extensively tested; existing studies quantify phonetic similarity using phonological features, rather than more direct acoustic or articulatory measures. In the second half of this paper (Section 3), I address this gap in the literature, by testing whether OCP-place in Samoan can be directly related to the perceptual similarity of consonants.

To summarize, the the current study aims to address the following questions:

1. Is reanalysis of Samoan thematic consonants sensitive to phonotactics, and if so, which phonotactic constraints are active?

2. If so, are the active phonotactic constraints phonetically natural, in the sense of having an articulatory or perceptual motivation?

Summing up, the role of phonotactics in morphophonology is still not well-understood. In this paper, I adopt a novel approach to examining this phonotactics/morphophonology connection, by looking at how similarity avoidance, specifically OCP-place, influences patterns of historical reanalysis. Additionally, I test whether OCP-place effects in Samoan correlate with measures of phonetic similarity; results will further our understanding of whether OCP-place has a functional phonetic motivation, and of whether learners are biased towards utilizing phonetically-motivated phonotactics.

## 2. Modeling reanalysis in Samoan

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 describe 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.[3]

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) | | |

[3]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.

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.

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

Table 2: Samoan thematic consonant alternations

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 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*).

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.[4] 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.

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'

[4]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.*qusa**n** 'to rain' is reflected as ua∼ua-i**n**a, rather than ua-**n**ia ), and /-na/ surfaces elsewhere. The /-ina/ here is homophonous with the vowel-initial /-ina/ allomorph.

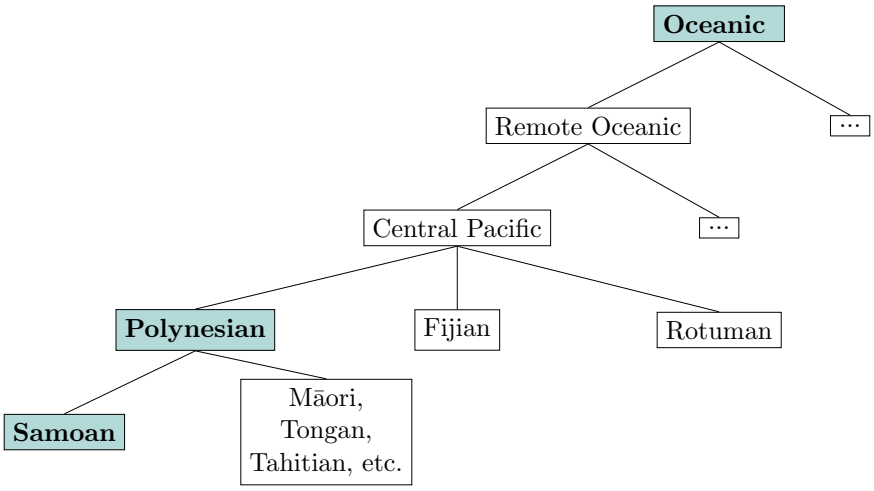



Figure 1: Internal subgrouping of Oceanic languages (Lynch et al., 2002; Pawley et al., 2007)

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/. A more detailed discussion of Samoan reanalysis, including discussion of the vowel-initial allomorphs, can be found in Kuo (2023b).

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 it is the most productive of all the suffixes that trigger thematic consonant alternations. The resulting wordlist has 593 stem-suffix pairs.

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%). If reanalysis is predictable from paradigm-internal frequencies, it should be primarily frequency-matching, with learners picking the new allomorph in a way that roughly matches the proportion at which they occurred in the lexicon. In other words, reanalysis should be in the direction of the vowel-initial allomorphs around 70% of the time, and towards other allomorphs

around 30% of the time. As discussed in Section 1.1, some degree of regularization towards more frequent variants might also be observed. As a result, the modern Samoan lexicon should mostly maintain the historical distribution, while also shifting towards the more frequent allomorphs.

Looking again at Fig. 2, we can observe that in modern Samoan, the proportion of vowel-initial allomorphs is roughly the same as it was in POc, with a slight increase (n=425/593, 72%). This is consistent with the predictions of a frequency-matching account. 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). Ergative forms that take /-ina/ are also sometimes omitted from the Milner (1966) dictionary. Nevertheless, the results suggest that reanalysis largely maintained the historical distribution of allomorphs, and was otherwise towards the more frequent vowel-initial variants.

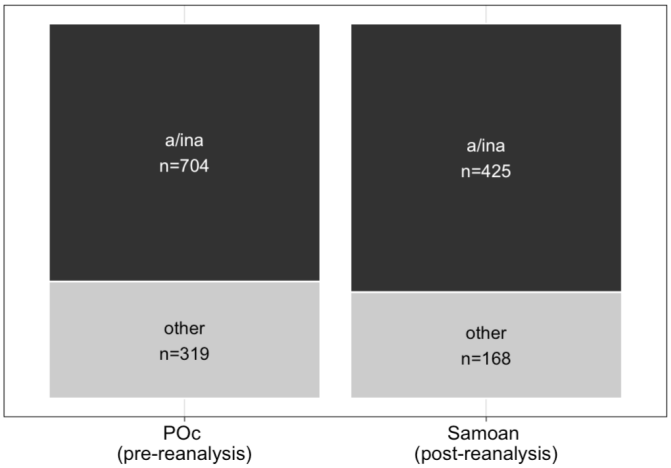



Figure 2: Distribution of ergative allomorphs before and after reanalysis

Table 4 shows a more direct comparison between POc and Samoan, and considers only the subset of items where both the Samoan form and POc protoform are available (n=147). These forms provide direct evidence for reanalyses that have occurred. For stems that were historically expected to take a vowel-initial allomorph, there has been very little reanalysis; only 8% of forms underwent reanalysis in the direction of /a, ina/→/Cia/ (where /Cia/ is any consonant-initial allomorph). In contrast, for the consonant-initial allomorphs, 53% of forms have been reanalyzed. Moreover, the majority of these reanalyses have been in the direction of /Cia/→/a, ina/ (towards the vowel-initial allomorphs).

| POc | Samoan (after reanalysis) | n | % |
|---|---|---|---|
| a, ina | unchanged | 66 | 0.92 |
| | to Cia | 6 | 0.08 |
| Cia | unchanged | 35 | 0.47 |
| | to a, ina | 31 | 0.42 |
| | to other Cia | 8 | 0.11 |

Table 4: Summary of reanalyses (POc protoforms vs. Samoan reflexes)

The data from Table 4 actually suggests that reanalysis has been towards the more frequent

variants, to a higher degree than would be expected if reanalysis was purely frequency-matching. This could be for reasons discussed in Section 1.1; for example, reanalysis could occur during earlier stages of language acquisition, resulting in overregularization. Additionally, however, reanalysis has also targeted certain /-Cia/ allomorphs at a higher rate than expected, for reasons I propose to be rooted in phonotactic considerations.

In particular, when we look at how the distribution of ergative allomorphs is conditioned by the identity of the preceding consonant, there are cases where reanalysis 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 occurred between POc and Samoan).

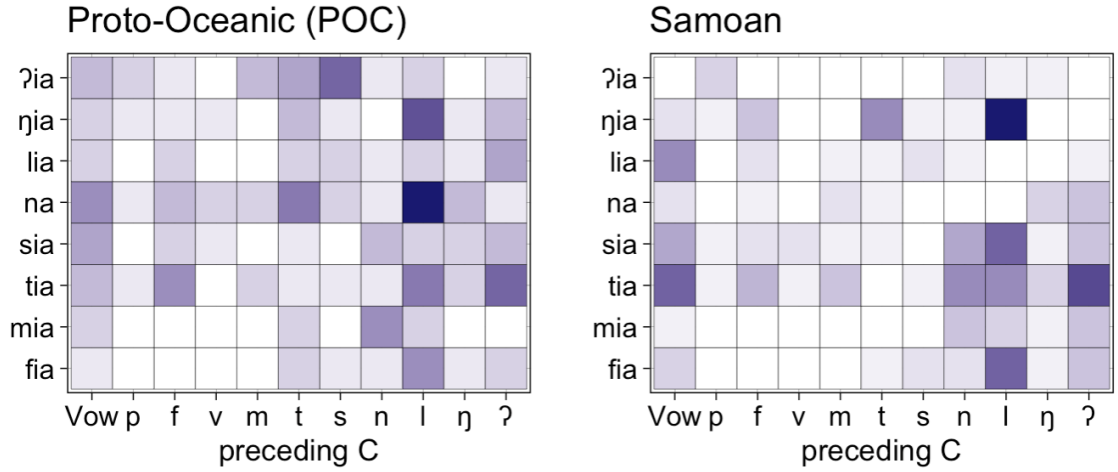



Figure 3: Distribution of ergative suffix allomorphs by preceding consonant in POc vs. Samoan

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 propose 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 gradient 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.

Note that in the POc data, there is an asymmetry between [l...n] sequences (which are well-attested) and [n...l] sequences (which are not well-attested). However, reanalyses has removed [l...n] sequences (i.e. suffixed forms of the type [i**l**o-n̲a]). This suggests that rather than learning

a specific restriction against [n...l] sequences, speakers were applying a more general constraint against sequences of (sonorant) 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).

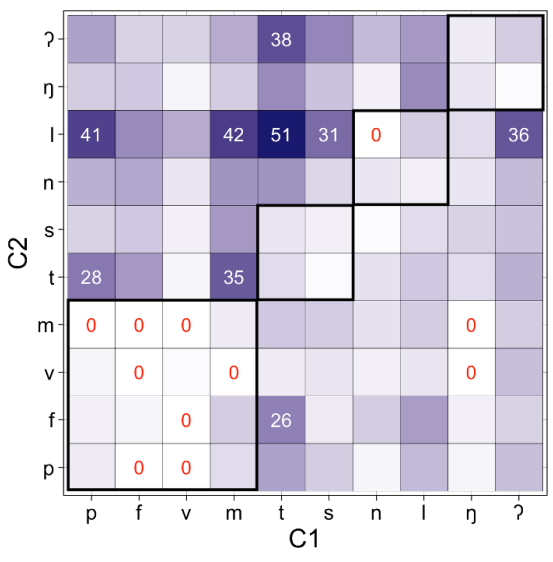



Figure 4: Consonant-consonant co-occurrences in Samoan

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 replicate the trends 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 is likely because [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 5 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, $\Delta$L 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 | $\Delta$L | 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** | **$7.56\times10^{-17}$*****|
| OCP-BACK | ŋaʔa, ʔaʔa,ŋaŋa | **3.94** | **0.002**** |
| OCP-BACK-SON | ŋaŋa, ʔaʔa | 0.01 | n.s. |

Table 5: OCP constraint weights learned by the phonotactic model

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.

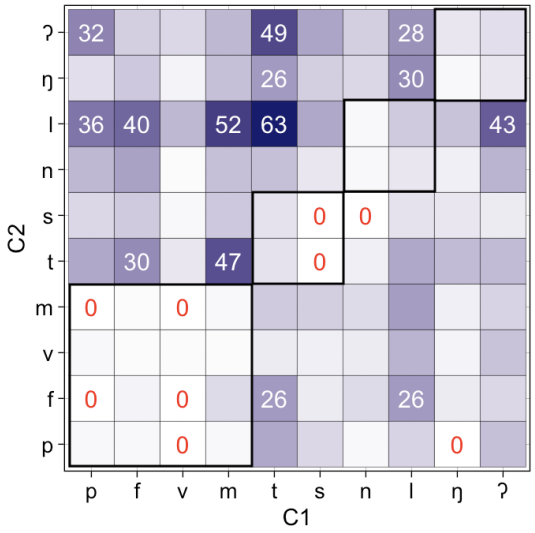



Figure 5: Consonant-consonant co-occurrences in PPn

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 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.

| CONSTRAINT | ΔL | P |
|---|---|---|
| OCP-place | 0.005 | n.s. (0.92) |
| OCP-LAB | **33.83** | $\mathbf{1.95\times10^{-16}}$*** |
| OCP-LAB-SON | 0.03 | n.s. (0.81) |
| OCP-COR | 0.99 | n.s (0.16) |
| OCP-COR-SON | **30.15** | $\mathbf{8.12\times10^{-15}}$ * ** |
| OCP-BACK | **3.97** | **0.005**** |
| OCP-BACK-SON | 0.01 | n.s. (0.36) |

Table 6: 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).[5]

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 [n] and the suffix allomorph is /-lia/.[6]

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)

[5]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.

[6]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, 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 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 ACTIVE 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.

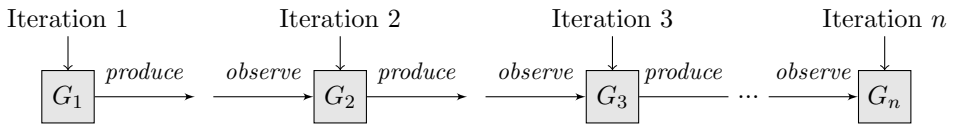



Figure 6: Structure of an iterated learning model, adapted from Ito and Feldman (2022, p. 3). $G_i$ indicates the learned grammar in each generation.

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 7 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 ($\Delta$L) 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 7: 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 difference 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.

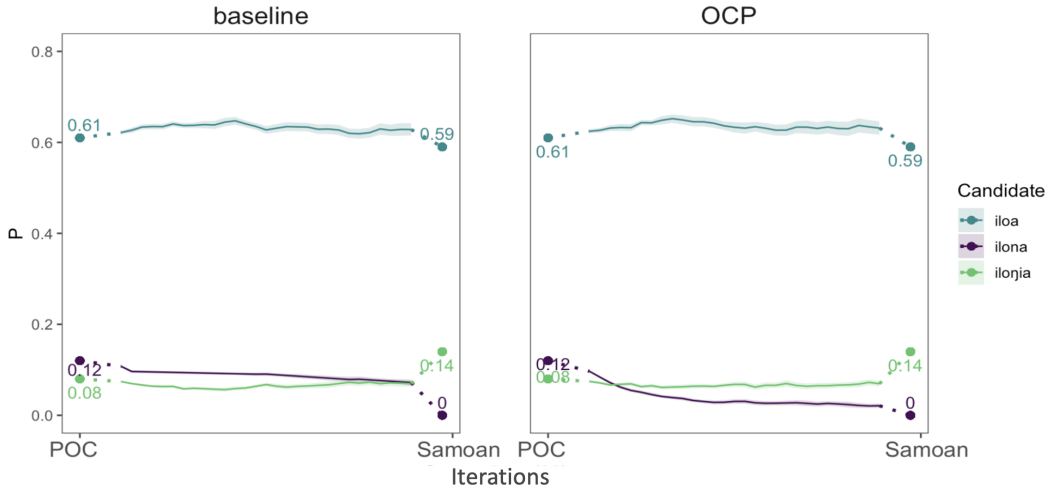



Figure 7: Model predictions in stems with a preceding /l/

*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 explore how both facts may be explained by the phonetic motivation behind OCP-place. Specifically, learners appear to selectively utilize phonotactics that are rooted in phonetic naturalness.

## 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 3.2-3.7 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 slow 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. Note that in most analyses of gradient OCP-place (e.g. Frisch et al., 2004; Coetzee and Pater, 2006; Wilson and Obdeyn, 2009), measures of consonant similarity are secondary to place of articulation. In other words, degree of similarity (or shared features) is only relevant for consonants that share the same place of articulation. For example, [n] and [m] might be globally more similar to each other than [n] and [l], but there are no restrictions against the co-occurrence of [n] and [m] because they do not share the same place of articulation.

Frisch (1996) and Frisch et al. (2004) 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). 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.

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

As an alternative, I propose that phonetic similarity can be quantified as the spectral distance between two phones, as described in Section 3.2. Note that spectral distance is a measure of perceptual similarity between sounds (to the extent that acoustic measures correspond to perception), so the results of this section shed light on how the *perceptual* similarity of segments can be directly related to the strength of OCP-place. As discussed in Section 3.7, however, similarity avoidance effects have been found in both production and perception studies, and although not the focus of the current paper, it is likely that OCP-place is also motivated by production-based similarity.

### *3.2. Method: measuring spectral distance*

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 coefficients

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).

To calculate MFCCs, the input signal is first transformed into a frequency domain using a technique like Discrete Fourier Transform. The resulting power spectrum is then resampled on the (log) mel scale, and a discrete cosine transformation is applied. The resulting set of MFCC coefficients contain information about how spectral energy is distributed across different frequencies.

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, many traditional phonetic measures, such as spectral peak, are only suited to measuring consonants with specific characteristics (in this case, ones with frication), and are not comparable across different manners of articulation. Other measures, such as consonant-vowel transitions, are comparable across most consonants, but primarily provide information on place of articulation and are less useful for capturing information about manner. Measures like formant tracking are also more susceptible to tracking errors.

In the current paper, MFCCs for consonants were obtained as follows. First, the speech signal was first segmented into phones (described in the following section). For each portion, 20ms of the preceding/following vowels were also included in the analysis, because formant transitions to/from surrounding vowels is a primary cue to consonant place. Because surrounding vowels are included in the analysis, tokens in phrase-initial position (with no preceding vowel) were excluded.

For each phone, the acoustic signal was downsampled to 12kHz. Each phone was blocked into frames of 25 ms duration, with a 10ms step. Each speech frame was parameterized into 13 coefficients, then for each token, the average MFCC across all frames was calculated. Following Gerosa et al. (2006), each averaged 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. Method: segmentation*

Consonants were manually aligned in Praat TextGrid (Boersma and Weenink, 2023) by a trained phonetician, using visual cues from the waveform and spectrogram. Representative examples of segmented phones are given in Fig. 8; the intervals labeled 'V' show the 20ms margins that were also included in the analysis of each token. Plosives (/p, t/) were marked from onset of the closure to end of aspiration; /t/ is consistently aspirated (VOT≈30-50ms), while /p/ typically has less aspiration and a weaker burst.

For consonants where the transition between surrounding vowels is less well-defined, vowel onset/offset was determined by the presence of clear and relatively steady formants in the vowels. For example, Fig. 9 shows the segmentation of [v], in a token where [v] has been lenited and has an approximant-like realization. Vowel onset/offset is marked at the point where both F1 and F2 are visibly darker and relatively stable.

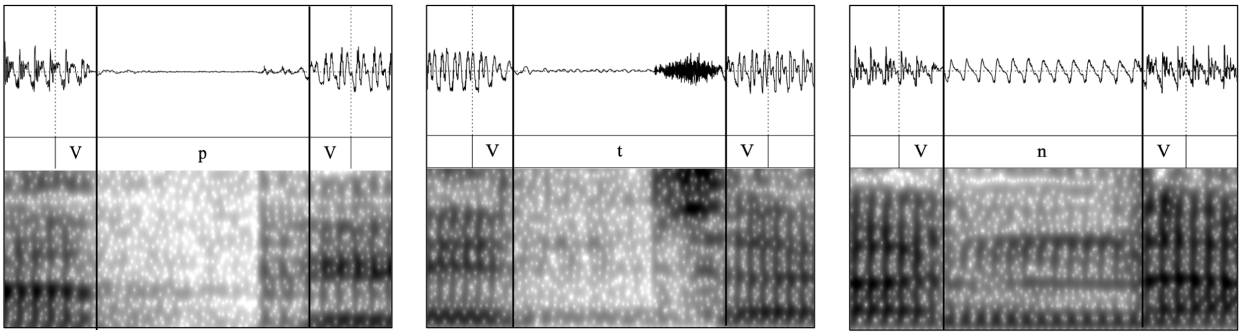



Figure 8: Examples of segmented consonants. Intervals labeled 'V' show the 20ms margin of vowels included in MFCC analysis.

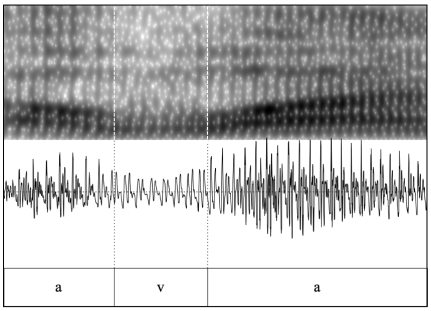



Figure 9: Segmentation of lenited [v]

### *3.4. Data*

Data comes from audio recordings of three male speakers from the Jehovah's Witnesses website.[7] 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; only three speakers are recorded with sentence-level transcription, limiting the number of speakers tested. Audio data is also 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.

In total, 1892 tokens were aligned and extracted; the distribution of tokens is summarized in Table 8. Note that the tokens are not evenly distributed across phonemes; this reflects the relative token frequency of each phoneme in Samoan.

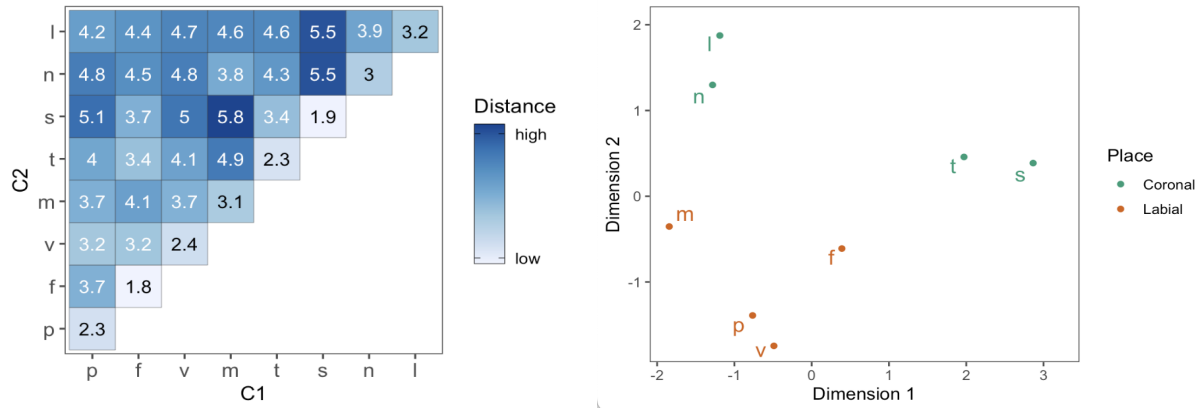
### *3.5. Results*

Fig. 10 visualizes the spectral distance between all target consonant-consonant pairs (i.e. all native phonemes, excluding [ʔ] and [ŋ]). Fig. 10a is a heatmap of the distances between each consonant-consonant pair, while Fig. 10b is a 2-dimensional visualization of the relative spectral distance between all consonants, obtained using Multidimensional Scaling (MDS, using the *cmdscale* function in R). MDS is a way to project the distance (or dissimilarities) between sets of objects (in this case, consonants) into a lower-dimensional and interpretable space. Consonants

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

| phone | N |
|---|---|
| p | 135 |
| f | 200 |
| v | 104 |
| m | 209 |
| s | 168 |
| t | 319 |
| l | 416 |
| n | 214 |

Table 8: Distribution of extracted tokens




(a) Mean spectral distance between consonant-consonant pairs (b) MDS visualization of the distance between consonants

Figure 10: Spectral distances across all target consonants

that are more similar (i.e. have shorter Euclidean distances) should be closer together on the graph than ones that are less similar.

Looking at these two figures, we see that overall, coronals and labials tend to fall into two separate clusters. However, segments which share the same place of articulation are not necessarily more similar to each other. For example, [n, s] are further apart (spectral distance = 5.5) than [n, m] (distance = 3.8). Additionally, the labials are more tightly clustered and closer to each other, suggesting that they are overall more acoustically more similar to each other. Within the coronals, there is a strong effect of manner; [l, n] and [t, s] each fall into two sub-clusters, but these two clusters are spaced far apart from each other.

Fig. 11 shows pairwise comparisons of spectral distance, grouped by place of articulation; 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. For ease of comparison, consonant-consonant pairs are also sorted in order of increasing spectral distance.

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.

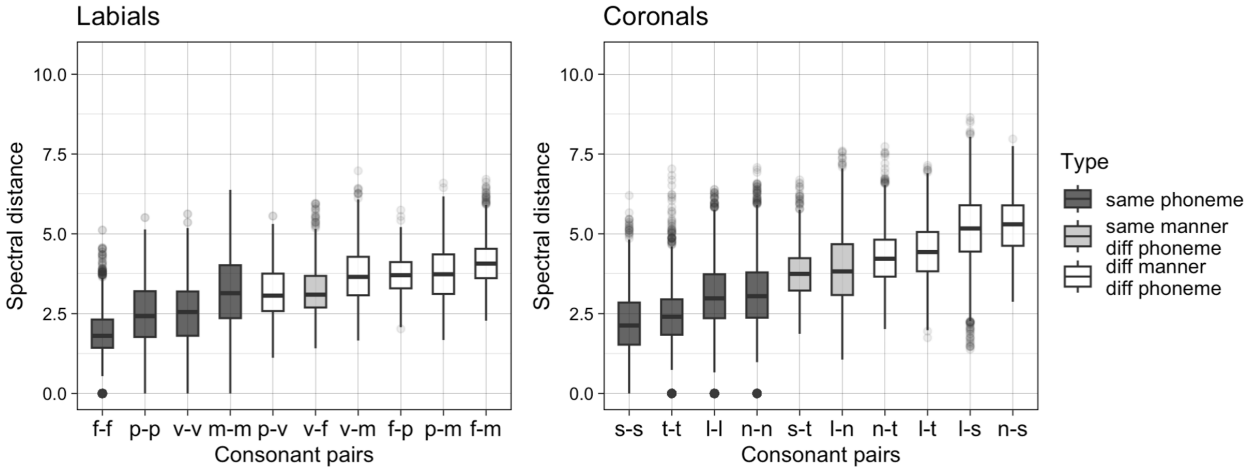



Figure 11: Spectral distances between consonant-consonant pairs, by place of articulation

A linear mixed effects regression (LMER) model was used to test if spectral distances vary by place of articulation. 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), PRECEDING VOWEL (baseline: /a/), FOLLOWING VOWEL, 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. Additionally, I tested for the interaction of PLACE and MATCH-SON, and the interaction of PLACE and MATCH-SEG.

Model results are summarized in Table 9; in the interest of space, results for PRECEDING VOWEL and FOLLOWING VOWEL are omitted; full model results can be found in Appendix 1. Using likelihood ratio testing (with the *anova()* function), all effects were found to be significant except for the interaction of PLACE and MATCH-SEGMENT. Consistent with figures above, increase in segment similarity (indicated by MATCH-SON and MATCH-SEG) results in a decrease in spectral distance. Additionally, there is an effect of PLACE, such that spectral distances are larger when the C1-C2 pair is coronal ($\beta$=0.86, CI = [0.82,0.91]). Interestingly, 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).

### *3.6. Relating spectral similarity to OCP-place restrictions*

Overall, spectral similarity appears to correspond closely to the OCP-place trends in the lexicon. In the phonotactics, as described in Section 2.3, there is a strong effect of OCP-LAB, which targets sequences of homorganic labials. However, there is not a strong effect of OCP-LAB-SON, which targets only labials that share the same sonorancy. In other words, for the labials, the similarity of segments does not strongly affect the strength of OCP-place; a sequence like [p...v] (both obstruents) is about as marked as a sequence like [p...m] (obstruent-sonorant). The opposite is true for coronals, where there is an active OCP-COR-SON constraint but not a general OCP-COR constraint. In other words, coronals strongly obey OCP-place *only* when they also share the same sonorancy. For example, a sequence like [n...l] (both sonorants) is marked, while [n...s] is relatively unmarked.

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 constrast, coronals

**Model**: dist ~ PLACE + MATCH-SON + MATCH-SEG + PREC-VOW + FOLL-VOW + PLACE:MATCH-SON + PLACE:MATCH-SEG + (1|speaker)

| *Predictors* | *Estimates* | *CI* | *p* |
|---|---|---|---|
| *Main effects* | | | |
| (Intercept) | 4.04 | [3.08, 4.28] | <0.001 |
| PLACE [cor] | 0.86 | [0.82, 0.91] | <0.001 |
| MATCH-SON [yes] | -0.57 | [-0.63, -0.50] | <0.001 |
| MATCH-SEGMENT [yes] | -0.98 | [-1.04, -0.93] | <0.001 |
| *Interaction effects* | | | |
| PLACE [cor] x MATCH-SON [yes] | -0.51 | [-0.58,-0.43] | <0.001 |
| PLACE [cor] x MATCH-SEGMENT [yes] | -0 | [-0.07,0.07] | *n.s.* |

Table 9: LMER model results for predictors of spectral distance between two segments

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.

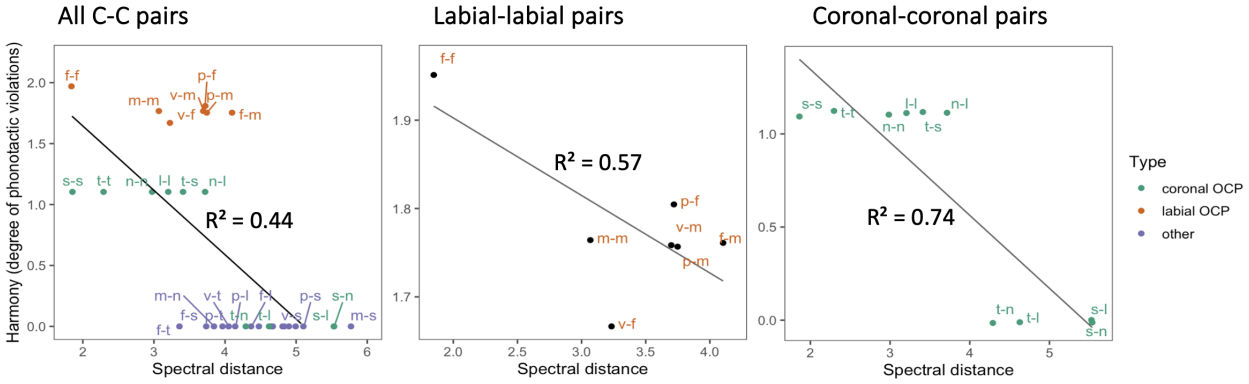



Figure 12: Spectral distance vs. phonotactic score (OCP grammar);figures show all C-C- pairs (left), the subset of labial-labial pairs (center), and the subset of coronal-coronal pairs (right)

To quantify this connection, the correlation between phonotactic scores and spectral similarity was evaluated. In Fig. 12, phonotactic scores from the OCP grammar in Section 2.4 are fit to spectral distance, for each consonant-consonant pair. The phonotactic scores are Harmony scores, which corresponds to the degree of phonotactic violation; the higher the harmony, the more phonotactically marked a sequence is.

Across all consonant-consonant pairs, there is a negative correlation between Harmony (degree of phonotactic violation) and spectral distance. In other words, the more spectrally similar two segments are, the more phonotactically marked they are. Note that there is a stronger correlation between spectral distance and harmony in the subset of forms which violate OCP-place (labial-labial pairs and coronal-coronal pairs); this suggests that similarity avoidance is primarily within each place of articulation, rather than across all consonant-consonant pairs. Additionally, the phonotactic grammar is more categorical than spectral distance; this is especially evidence for the coronal-coronal pairs, which fall into two distinct groups in their phonotactic violation profile, but

have a gradient range of spectral distance.

The fit between phonotactics and spectral distance is is weaker for the labials, likely because the harmony scores for all labial-labial sequences are high, with very little variation (1.67-1.97). For the coronals, there is a strong correlation ($R^2$=0.74). In other words, coronals overall vary more in their spectral distance, in a way that is consistent with the effect of sonorancy in the phonotactics.

### *3.7. Discussion*

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.

In this study, phonetic similarity was quantified using a measure of spectral distance. This is novel compared to existing work on gradient OCP-place, which typically uses featural similarity as a proxy for phonetic similarity. In Samoan, and crosslinguistically in languages where OCP-place is active, OCP-place tends to be stronger in labials than in coronals. In the current study, this falls out from the fact that labials are more similar than coronals in terms of MFCCs.

As pointed out by a reviewer, in models based on natural classes, including the one implemented by Frisch, Pierrehumbert and Broe (2004), the different patterning of labials and coronals would actually fall out from the fact that most languages have larger sets of coronal than labial consonants. Consequently, there are more natural classes within the coronals, and that results in coronals being generally less similar to each other.

This approach would not work straightforwardly for Samoan, which has an equal number of (non-loan) labial and coronal phonemes (/p f v m/ vs. /t s n l/). Instead, in Samoan, the different patterning of labials and coronals is more likely the result of language-specific details about phonetic realization. For the plosives, Samoan [p] is unaspirated (or has short-lag VOT), while [t] is consistently aspirated, and in this sense has stronger perceptual cues to distinguish it from other manners of articulation. In the fricatives, [s] is a sibilant and therefore likely more perceptually salient.

In Samoan, a phonetic similarity account of OCP-place outperforms a featural similarity account. If this phonetic similarity account is to generalize beyond Samoan, however, there still needs to be an explanation for why crosslinguistically, OCP-place tends to be stronger for labials than coronals. Though a full exploration of the issue is beyond the scope of this paper, I offer speculations on two possible mechanisms. First, this asymmetry may be rooted in intrinsic acoustic differences between labials and coronals. Coronal obstruent bursts could be more perceptually salient because they have a spectrum that is shaped by cavity resonances, compared to labials which have a more diffuse burst. Prior studies, such as Chang et al. (2001), have found that changes to coronal obstruents are more perceptible because of such acoustic differences.

Another, alternate account is rooted in dispersion and contrast maximization (Lindblom and Maddieson, 1988; Schwartz et al., 1997; Flemming, 2004): the more categories a language contrasts, the more the language should make use of a phonetic space. Because coronal inventories are often larger than labial inventories, dispersion-based accounts would predict that they make greater use of the phonetic space. As a result, coronals are phonetically less similar to each other, and less restricted by OCP-place. In general, more careful crosslinguistic work, focused on languages where spectral distance and featural similarity might make different predictions, is needed to tease apart featural and phonetic similarity.

## 4. General discussion

Models of reanalysis (and more generally of morphophonological learning) typically focus on the effect of paradigm-internal frequencies. In the present study, I explored the possibility that reanalysis is also sensitive to phonotactics, but in a way that is constrained by phonetic naturalness. Results of Section 2 support an analysis where phonotactics can influence reanalysis. In particular, in modeling Samoan reanalysis, I find that a model which incorporates phonotactics outperforms one that only uses paradigm-internal frequencies.

Importantly, the choice of phonotactics also appears to be constrained; Samoan has a phonotactic constraint against sequences of homorganic consonants, characterized in Optimality Theory as OCP-place. In Section 2, 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 preferentially utilizes phonotactics that have phonetic motivation. OCP-place, specifically, is rooted in general processing constraints against the repetition of phonetically similar segments. The acoustic study in Section 3 supports this analysis, and finds close correlation between OCP-place phonotactics and the spectral distance between consonants.

More generally, results of Section 2 show that in studying morphophonological learning, the effect of phonotactics should be considered. Additionally, the phonetic basis of phonotactic constraints also matters, as learners may preferentially utilize constraints that are phonetically motivated. Notably, effects of phonotactics and phonetic substance in morphophonology can be difficult to identify. 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, learners have been shown to 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.

Results of Section 3 also support Frisch et al.'s (2004) 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 the acoustic study in Section 3, I find similar support that OCP-place is rooted in phonetic similarity avoidance. 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.

Findings from Section 3 also suggest that perceptual similarity avoidance in Samoan (as quantified by spectral distance) only affects OCP-place within each place of articulation. When all consonant-consonant pairs are considered (as in Fig. 10), some consonant pairs are acoustically very similar to each other, but are not sensitive to any co-occurrence restrictions in Samoan (e.g. [n, m]). In these cases, the consonants do not share the same place of articulation. This is consistent with the existing literature on OCP-place, which shows that similarity (typically measured in terms of featural similarity) is secondary to place of articulation (for a review, see Wilson and

Obdeyn, 2009). That is, similarity modulates the strength of OCP-place effects *within* consonants that already share the same place of articulation.

Prior literature suggests that similarity avoidance is rooted in both production (e.g. Dell, 1984; Sevald and Dell, 1994) and perception (e.g. Miller and MacKay, 1994). The focus of this study was perceptual similarity, approximated using spectral distance. Although spectral similarity by itself appears to be strongly predictive of the strength of OCP-place, the Samoan results are still compatible with an analysis where gradient OCP-place is motivated by a mixture of production-based and perception-based similarity. In Samoan, consonant co-occurrence restrictions primarily target sounds that share the same place of articulation, and this could be rooted in avoidance of articulatorily similar sounds. Within each place of articulation, perceptual similarity then modulates the strength of the OCP-place constraint. Existing work on gradient OCP-place does not always make a clear distinction between the effects of perceptual and articulatory similarity. Going forward, future work should focus on teasing apart the effects of different types of phonetic similarity.

## 5. Conclusion

In this paper, I find 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. These results provide new evidence that fine-grained phonetic detail can influence the learning of morphophonological paradigms.

In Section 2, patterns of reanalysis in Samoan stem-ergative pairs were examined and modeled. Results suggest that reanalysis is generally frequency-matching, as it tends to be in the direction of the vowel-initial allomorphs /a, ina/, which were historically the most frequent. However, reanalysis has targeted suffixed forms like [ila-na] more than would be expected in a frequency-matching approach. I propose that this is the effect of OCP-place, which is a phonotactic constraint against sequences of homorganic consonants. For example, [ila-na] contains a sequence of coronal sonorants [l...n], making it phonotactically marked and prone to reanalysis.

OCP-place is argued to have a functional phonetic motivation, in that it is the avoidance of phonetically similar sequences of consonants. However, in existing work that quantifies segment similarity in OCP-place, phonological features are typically used as a proxy for phonetic similarity. In this paper, I instead use a more direct measure of phonetic similarity, by measuring the distance between the MFCC vectors of target segments. Results show that there is a correlation between the strength of OCP-place and the spectral similarity of segments, supporting the characterization of OCP-place as phonetic similarity avoidance.

## Acknowledgements

I would like to thank the editor as well as three anonymous referees for their thorough and helpful feedback; this article benefited greatly from their input. Additionally, I would like to thank Bruce Hayes, Kie Zuraw, Claire-Moore Cantwell, David Goldstein, and the audiences at HISPhonCog 2023.

## 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.

Blevins, J., 2008. Consonant epenthesis: natural and unnatural histories, in: Good, J. (Ed.), Language universals and language change. Oxford University Press, pp. 79–107.

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.

Chang, S.S., Plauché, M.C., Ohala, J.J., 2001. Markedness and consonant confusion asymmetries, in: Johnson, K., Hume, E. (Eds.), The role of speech perception in phonology. Brill, pp. 79–101.

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.
Flemming, E., 2004. Contrast and perceptual distinctiveness, in: Hayes, B., Steriade, D., Kirchner, R. (Eds.), Phonetically based phonology. Cambridge University Press, pp. 232–276.
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.
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.
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.
Lindblom, B., Maddieson, I., 1988. Phonetic universals in consonant systems, in: Hyman, L.M., Li, C.N. (Eds.), Language, speech and mind. Routledge London, pp. 62–78.
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.
Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., Xu, F., Clahsen, H., 1992. Overregularization in Language Acquisition. volume 57 of *Monographs of the Society for Research in Child Development*. URL: https://www.jstor.org/stable/1166115?origin=crossref, doi:10.2307/1166115.
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.
Schumacher, R.A., Pierrehumbert, J.B., 2021. Familiarity, consistency, and systematizing in morphology. Cognition 212, 104512.
Schwartz, J.L., Boë, L.J., Vallée, N., Abry, C., 1997. The Dispersion-Focalization Theory of vowel systems. Journal of Phonetics 25, 255–286. URL: https://linkinghub.elsevier.com/retrieve/pii/S0095447097900437, doi:10.1006/jpho.1997.0043.
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.

## Appendix 1

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

| *Predictors* | *Estimates* | *CI* | *p* |
|---|---|---|---|
| *Main effects* | | | |
| (Intercept) | 4.04 | [3.08, 4.28] | <0.001 |
| PLACE [cor] | 0.86 | [0.82, 0.91] | <0.001 |
| MATCH-SON [yes] | -0.57 | [-0.63, -0.50] | <0.001 |
| MATCH-SEGMENT [yes] | -0.98 | [-1.04, -0.93] | <0.001 |
| PRECEDING-VOWEL | | | |
| [e] | -0.38 | [-0.42, -0.35] | <0.001 |
| [i] | -0.25 | [-0.29, -0.20] | <0.001 |
| [o] | 0.17 | [0.13, 0.21] | <0.001 |
| [u] | -0.26 | [-0.34, -0.18] | <0.001 |
| FOLLOWING-VOWEL | | | |
| [e] | 0.08 | [0.05, 0.12] | <0.001 |
| [i] | -0.44 | [-0.50, -0.38] | <0.001 |
| [o] | -0.33 | [-0.47, -0.35] | <0.001 |
| [u] | -0.51 | [-0.58, -0.43] | <0.001 |
| *Interaction effects* | | | |
| PLACE [cor] x MATCH-SON [yes] | -0.51 | [-0.58,-0.43] | <0.001 |
| PLACE [cor] x MATCH-SEGMENT [yes] | -0 | [-0.07,0.07] | *n.s.* |

Table .10: LMER model results for predictors of spectral distance between two segments