To appear in Linguistic Discovery.

Samoan root phonotactics: Digging deeper into the data*

John Alderete, Mark Bradshaw

Simon Fraser University

Abstract. This article gives a detailed quantitative account of Samoan root phonotactics. In particular, count data is given in eleven tables of segment frequencies (i.e., consonants, short and long vowels, diphthongs) and frequencies of combinations of segments (i.e., syllable types, consonant-vowel combinations, V-V and C-C combinations across syllables). Systematic patterns of over- and under-representation of these structures in the lexicon are documented and related to prior research. Beyond the detailed frequency facts presented here, new empirical patterns documented include positional preferences for bilabials and non-labial sonorants, extensions of a known pattern of gradient vowel assimilation, and identification of a role for manner and segment order in consonant co-occurrence restrictions.

Keywords: Samoan, phonology, phonotactics, gradience, dissimilation, assimilation, frequency

1. Introduction

The study of phonotactics, or the system of legal sound combinations, in Austronesian languages has led to many important scientific findings. In the domain of phonetics, Maddieson & Precoda (1992) examine the lexical statistics of C-V combinations in Hawaiian and use it as a basis for evaluating phonetic theories of articulatory ease and acoustic salience. In phonology, phonotactics has been used as a basis of classifying unusual segments (Francois 2010), proposing new theories of dissimilation and root co-occurrence restrictions (Berg 1989, Coetzee & Pater 2008, Harlow 1991, Mester 1986, Uhlenbeck 1950), and documenting language-internal pressures that motivate phonological processes (Blust 2007). Recent work on Austronesian languages has even shown how phonotactics impacts morphology, as in how exceptions to the well-known phonotactic pattern of nasal substitution in Tagalog guides novel word productions (Zuraw 2000), or how the phonotactic probabilities of stem-final consonants can be used as a basis for predicting morphologically derived forms in Maori (Jones 2008).

Each of these studies depends crucially on detailed descriptions of sound structure frequencies. This reliance on frequency underscores the importance of quantitative analysis in describing phonological structure and the research above attests to its importance in the burgeoning field of probabilistic linguistics.

This article provides a detailed quantitative account of phonotactics in Samoan roots, an account that is guided and motivated by prior research. Mosel & Hovdhaugen (1992), an authoritative reference grammar of Samoan, documents many important facts of segment occurrence in Samoan, including qualitative statements about low and high frequency segments and restrictions on combinations of labial consonants. These restrictions contain many exceptions referring to specific segments that we examine in detail below. Another important work, Krupa (1967), investigates consonant-consonant and vowel-vowel co-occurrence across syllables in Samoan,

*Thanks to Paul Tupper, Lindsay Whaley, and two anonymous reviews for his comments on an earlier draft of this article. This work was funded in part by SSHRC standard research grant 410-2005-1175.

with the ultimate goal of clarifying the structural similarities among Polynesian languages for use in linguistic classification (see also Krupa 1966, 1968, 1971). The principal findings are that Samoan, like many other Polynesian languages, has gradient patterns of vowel assimilation in non-low vowels and a systematic avoidance of homorganic consonants in roots. While sufficiently detailed for Krupa's larger study of linguistic classification, the organization of natural classes in his work did not identify certain factors that have since been shown to be important in characterizing consonant co-occurrence, including the distinction between identical and non-identical consonants, manner, and segment order. The present work builds on Krupa's findings by documenting segment phonotactics with these and other factors in mind.

The larger goal of this work is thus to 'dig deeper' into the data and provide a more comprehensive account of the frequency structure and relationships among sound structures in Samoan phonology. We see the documentation of this frequency structure as an important part of the description of Samoan sound structure and worth documenting for its own sake. In addition, we show how these facts relate to phonological and phonetic theories of syllables and phonological constraints across syllables. Our goal is not to analyze each new empirical finding, but rather clarify some important descriptive patterns of relevance to contemporary phonology such that future phonological analysis can stand on stronger empirical ground.

The rest of this article is organized as follows. Section 2 provides background on Samoan phonology, linguistic history, and morphology necessary to understanding the phonotactic generalizations. The next section describes the methods used in constructing the root lists, the primary source of data for the quantitative analyses. Section 4 documents the frequencies of segments in isolation and frequencies of syllable patterns, some of which are implicated in sections 5 and 6, where vowel-vowel and consonant-consonant co-occurrence are explored in more detail. The last section summarizes the principal findings and sketches some of the implications they have for further analysis in generative phonology.

2. Background

Samoan is a member of Nuclear Polynesian because it has a set of shared phonological (Elbert 1953; Mosel & Hovdhaugen 1992) and morphological (Pawley 1966) innovations that exclude it from Tongic (i.e., Tongan and Niuean). Within this group, Samoan is considered part of Samoic Outlier, coordinate with Eastern Polynesian, because it has a set of morphological and morphophonemic features that are exclusive of Tongic and Eastern Polynesian (Pawley 1966). That said, Samoan has many of the phonological features typical of Polynesian languages, e.g., a small phoneme inventory, five vowels that contrast in length, and exclusively open syllables (Krupa 1973, Blust 2009), as can be seen from the sound inventory below.

(1)

Consonants				Vowels			
p	t	ʔ		i	u	iː	uː
f v	s			e	o	eː	oː
m	n	ŋ		a		aː	
	l						

Compared to other Polynesian languages, Samoan is relatively conservative in retaining most sounds from Proto-Polynesian. Exceptions to this statement in Elbert's (1953) historical analysis

include the loss of glottals **h *ʔ* (Samoan *ʔ* is a reflex of **k*), the merging of **l* and **r* to *l*, and certain vowel mergers. Though not treated as such by Elbert, Samoan *v* is the historical reflex of **w*, as shown by reconstructions of most *v*-initial words on Pollex-Online (http://pollex.org.nz/). While *k, r* and *h* were all lost, they are used in loan words, with *h* occurring word-initially in most cases, e.g., *haikomisi* 'high comission'.

There are two sociolinguistic levels of Samoan, Tautala lelei (literary language) and Tautala leaga (colloquial), and native speakers are generally cognizant of both levels. Our description of the phonotactics focuses on the literary language, following most prior work, but the principal difference between the two levels can be summed up with three phoneme mergers in Tautala leaga: *t* and *k* > *k*, *n* and *ŋ* > *ŋ*, *r* (from loans) and *l* > *l* (Mosel & Hovdhaugen 1992).

Syllables in Samoan generally conform to the template, $(C)V_1(V_2)$, where coda consonants are forbidden but simplex onsets and the second of two vowels are optional. Long vowels are assumed here to have a single vowel quality with two moras, and diphthongs, also bimoraic, can be composed of any non-high vowel plus a high vowel, or *ui*, following the analysis of diphthongs in (Mosel & Hovdhaugen 1992). Thus, the following VV sequences are assumed to be tautosyllabic morpheme-internally: *ei eu ai au ou oi* and *ui*. All others are treated as heterosyllabic. Stress falls on the final syllable if it contains a bimoraic nucleus and otherwise on the penult. Evidence for the tautosyllabic syllabification of the above VV sequences comes from the fact that, in normal speech, stress falls on the first of two vowel structures in the penult, e.g., *téi.ne* 'girl', *táu.a* 'war', cf., *va.ó.a* 'be overgrown with weeds', in which neither *ao* or *oa* can form a tautosyllabic diphthong. There are special contexts, however, in which a strict penultimate mora pattern is found, e.g., between morpheme boundaries */fe+ita/* → *feíta* 'be angry (pl)' and certain ideolectical speech patterns (see Mosel & Hovdhaugen 1992: 30-31).[1]

Some brief remarks on the morphology of nouns and verbs will also help the discussions below. Samoan morphology is largely derivational—there are no inflectional paradigms in the traditional sense. The main derivational categories marked by the morphology are causatives, plurality of subject/events/patients, frequentative, reciprocal, ornative, (de-)ergative, conversion of noun to verb, and vice versa. Most derivational categories can be marked by more than one process, and conversely, certain morphological processes, like partial reduplication, can mark a host of distinct semantic categories. The most common and productive prefixes are: *fa?a-* (causative), *fe-* (plurality of events, reciprocal), and *ma-* (de-ergativizer). Common suffixes include: *-e* (vocative), *-(in)a* (ergativizing), *-a* (ornative), *-(C)ia* (ergativising), *-ŋa* (nominalizer). In addition to compounding, Samoan has several patterns of reduplication and a nonproductive nonconcatenative process marked by vowel lengthening.

[1] The syllabification of VV sequences in Samoan has some unresolved issues, including the analysis of the special circumstances in which diphthongs are heterosyllabic mentioned above, and even the tautosyllabic status of a sequence of identical vowels. Mosel & Hovdhaugen (1992), for example, argue on the basis of the gliding of long mid vowels and partial reduplication that certain consecutive identical vowels are heterosyllabic. After reviewing this evidence, we do not see how these patterns are inconsistent with the tautosyllabic analysis assumed here, which is our best attempt to integrate a wide range of facts. We acknowledge, however, that there are unresolved issues.

3. Composition of the root list

The root list that constitutes the primary data source for this work is a list composed of all the noun and verb roots from Milner (1966). The root list is available from the first author's webpage as a data supplement, together with tabular data that is only summarized here. The root list is of course only a sample of the true population of Samoan roots, because Milner's dictionary is also a sample of this population. However, it is comprehensive in the sense that it involved extracting all roots of content words from the dictionary, and the dictionary attempted to list all roots. The rest of this section summarizes the assumptions used to extract roots from the dictionary, which are explained in more detail in the README file associated with the data supplement.

The root list contains the headwords of free content morphemes in the Milner dictionary, i.e., unbound nouns and verbs, and a handful of adjectives. Loanwords were excluded because they are known to exhibit non-native phonological patterns. It is important to note that we extracted headwords only. This assumption had the effect of excluding noun or verb roots that are bundled in the same entry with a related root, which is not uncommon in the Milner dictionary. Also, because the phonological descriptions below are intended to document type frequencies, we included homophones that had nonobvious meaning differences. For example, both *fao* 'nail (n)' and *fao* 'small shore tree (n)' were included.

We used canonical root shape criteria to ensure included items were in fact simplex roots and not morphologically complex. In particular, we assumed that roots were no more than four moras (Krupa 1966, Krupa 1971), and further scrutinized the entry of each headword for evidence of complexity, including common affixes, reduplication, and cross-referencing. Bound morphemes, shown in parentheticals in the dictionary, were also excluded. Entries for classificatory names of animals, plants, seafood, etc., which are all nouns, were also closely examined because these names can be rather idiosyncratic and difficult to find evidence of complex structure. Thus, all one or two syllable classificatory names were taken without hesitation. All others were included in the larger comprehensive spreadsheet (because they may be relevant for other study), but removed from the principal analyses given below because of the difficulty in assessing morphological complexity.

The total number of roots in the comprehensive root list, including classificatory names greater than two syllables, is 1871. Excluding just three and four syllable classificatory names, and adjectival roots, reduces this list to 1640, which is the main root list used in analysis below. In some cases, to compare our findings with those of prior work, we limit our examination further to just disyllabic noun and verb roots, which is a list of 1137 roots.

4. Descriptive statistics of segment and syllable patterns

It is common to find qualitative statements in reference grammars about how common or rare a segment is, and some grammars even give quantitative analysis of this type of fact. In Samoan, we find some statements of the former kind in Mosel & Hovdhaugen (1992), for example, a phonotactic restriction against *ŋ* is conjectured to be due to its low frequency. However, the frequencies of Samoan sounds and CV sequences have never been given a rigorous and comprehensive treatment. We therefore begin our study of Samoan phonotactics with some descriptive statistics of basic segment and syllable frequencies.

A first pass over the main list of 1640 roots gives a gross sense of the size of root words and the relative frequency of basic syllabic types. As shown below, there are very few four syllable roots, and monosyllabic roots are also rather rare. Disyllabic roots constitute the largest percentage of the root list (69%), followed by trisyllabic roots (21%). As for syllable types, CV syllables (with short vowels only) are by far the most common type in all contexts (69% of all 3644 observations), followed by CVV (16%) and _V (13%), which in turn is trailed by onsetless syllables with long vowels _VV (2%). These frequencies seem to follow well-known markedness scales, CV < _V, CV < CVV, where marked structures, like onsetless syllables and long vowels, occur with lower frequency (see e.g., Blevins (1995) for justification of these markedness assumptions).

Table 1. Word size frequencies

1	2	3	4	Total
147	1137	345	11	**1640**

Table 2. Syllable type frequencies

	Initial	Medial	Final	Total
CV	1096	295	1136	**2527**
CVV	284	3	278	**565**
V	210	69	183	**462**
VV	48	0	42	**90**

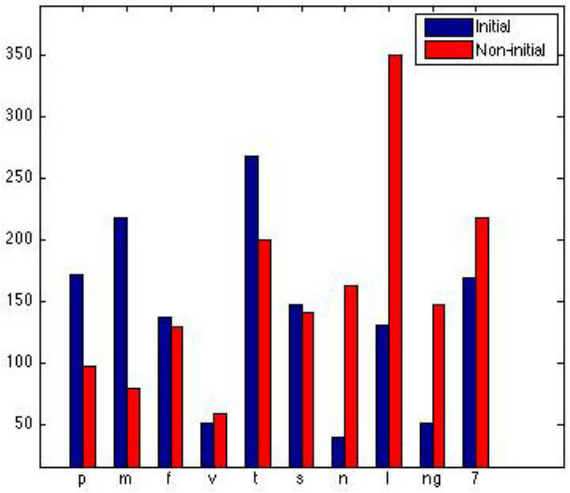
Medial syllables, i.e., non-initial and non-final syllables in tri- and quadrasyllabic roots, almost exclusively contain short vowels. However, there is no apparent preference for any syllable type to appear initially or finally.

Moving next to segment frequencies, the table below gives the frequency of each consonant in either the initial or non-initial syllable, and overall frequencies are shown in the last column. From these numbers, it can be gleaned that the most frequent consonants are *t, l,* and *ʔ,* constituting roughly 45% of all observations. The next most frequent consonants are *m, s,* and *f*, trailed by *n* and *ŋ,* and finally *v*, which is the lowest frequency consonant.

Table 3. Consonant frequencies

	initial	non-init	Total
p	172	97	**269**
m	218	79	**297**
f	137	129	**266**
v	51	58	**109**
t	268	200	**468**
s	147	141	**288**
n	39	162	**201**
l	130	350	**480**
ŋ	51	147	**198**
ʔ	169	218	**387**
Total	**1,382**	**1,581**	**2,963**

Figure 1. Consonant frequencies by position



An interesting set of distributional facts is apparent from the data above, illustrated in the adjacent figure. It has been noted in prior work that labial consonants have a greater than

chance occurrence in initial syllables in many related Oceanic languages (Krupa 1966). However, in this work, the restriction is a general statement about all labials, including Krupa's (1967: 73) statement to this effect for Samoan. The data above for Samoan reveal that specifically bilabials *p* and *m* are preferred in initial position, by a greater than 2-to-1 margin, but labial-dentals *f* and *v* have no such preference. On the other hand, non-labial sonorants, i.e., *n, l* and *ŋ*, are much more common in non-initial syllables. The comparison among the nasals is quite striking: while only 26.6% of all *m* occurrences are non-initial, *n* and *ŋ* occur non-initially 80.6% and 74.2% of the time, respectively. Note that this skewed distribution cannot be due to the fact that there are more non-initial positions than initial positions. As show in Table 2, the difference between initial and non-initial observations is only 332, so the two positions are split roughly 45% (initial) vs 55% (non-initial). It will be shown in section 6 that this positional preference appears to have an effect on the restriction against homorganic C-C's across syllables, because most of the exceptions to this restriction involve initial bilabials and non-initial coronal sonorants.

Vowel frequencies are shown below, again sorted by initial and non-initial positions. From these tables, we see that *a* is by far the most common short vowel, while *e* is the least frequent, and the other three vowels have roughly equal frequency. With the exception of *a*, there does not seem to be a preference for position within a word. These trends are repeated in the long vowels, which are of course far less frequent than short vowels.

Table 4. Short vowel frequencies

	initial	non-initial	Total
i	168	230	**398**
e	116	174	**290**
a	591	242	**833**
o	209	211	**420**
u	234	148	**382**

Table 5. Long vowel frequencies

	initial	non-initial	Total
iː	22	6	**28**
eː	14	8	**22**
aː	112	42	**154**
oː	29	17	**46**
uː	26	22	**48**

Relative to short vowels, diphthongs are rather rare. The table below reports the frequencies of diphthongs, excluding diphthongs in which one member is a long vowel. Recall that diphthongs always contain a high vowel as a second component, and *iu* is not a permissible diphthong (see section 2).

Table 6. Diphthong frequencies

	i	u	Total
i	*NA*	0	**0**
e	16	7	**23**
a	45	83	**128**
o	16	15	**31**
u	17	*NA*	**17**

The data above show that diphthongs starting with *a* are by far the most common. We also note that 7 of the 17 instances of *ui* are in fact word initial, as in *ui?i* 'youngest of a family'.

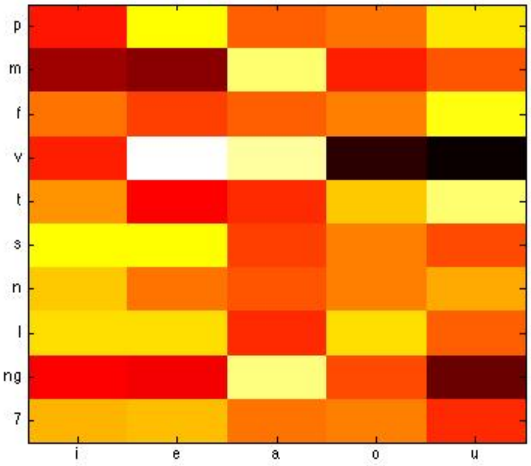
Mosel and Hovdhaugen (1992) state (p. 29) that this sequence is pronounced with an initial glide *w* initially, and so these cases may not be properly understood as phonetic diphthongs.

The next table combines the consonant and vowel structures discussed above and gives a count of their combinations in the main syllable type in Samoan, CV. The frequencies below include counts of both short and long vowels, and also CV sequences in which the V is the first member of a diphthong. The figure to the right is a visualization using a heat map of the adjusted frequency data, mapping it onto a color scale in which lighter colors indicate a greater relative frequency of a given CV combination, and darker colors, like the black boxes for *vo* and *vu*, indicate lower frequency relative to the rest of the data. This adjusted frequency is the ratio of Observed/Expected (O/E), or the number of observed CVs over the number of CVs of this type that would be expected if they were to occur at chance, a common way of standardizing type frequency data in phonotactics (see e.g., Pierrehumbert 1993). A structure with an O/E value far below 1.0 is standardly interpreted as being under-represented in the lexicon, and conversely, an O/E value far above 1.0 indicates over-representation. (Actual O/E data for CV frequency is provided in the online data supplement.)

Table 7. CV frequencies

	i	e	a	o	u
p	31	40	101	43	54
m	20	12	188	36	42
f	42	25	98	44	57
v	13	22	72	2	0
t	79	35	147	93	113
s	62	43	96	47	39
n	40	23	76	35	37
l	96	67	148	98	68
ŋ	21	14	127	28	8
ʔ	71	51	155	63	47

Figure 2. Heat map of CV O/E



The only clear gap here is the nearly categorical exclusion of *v* before back vowels, a common finding for *w* in other languages, and Samoan *v* is in fact the historical reflex of **w*. The larger finding in this dataset, however, is that Cs and Vs seem to combine more or less randomly, a fact consistent with Maddieson & Precoda's (1992) finding for a host of other languages. This work examined CV frequencies in five languages with a focus on testing the role of principles of adaptive dispersion and quantal change in shaping lexicons. One of the findings of this work was that major place of articulation of an onset consonant did not have a strong effect on the identity of the following vowel in the five languages examined. This was measured with a set of vowel deviation scores that indicated how much a CV syllable beginning with a particular place class deviated from the norm. An example helps show how these vowel deviation scores are calculated. Suppose that the percentage of syllables beginning with *pi*, P_{pi}, is 11.5% of all syllables that began with *p* and are followed by a vowel. Also, the percentage of *i* occurrences out of all vowel occurrences, $=P_i$, is 15.11%. Thus, the deviation of the former from the latter is

P_i-P_{pi}, or 3.59. Low deviation scores for many segments in the same class supports a conclusion that the class of the consonant is not a strong factor in predicting CV frequencies.

Following this same procedure, the vowel deviation scores are given for all labials and coronals below, shown as a mean of four observations (other place classes are not included because there is only one velar and one glottal). The vowel deviation scores for Hawaiian syllables are given for comparison, drawn from Maddieson & Precoda (1992).

Table 8. Vowel deviation scores across labials and coronals

Labials	i	e	a	o	u
Samoan	3.6	-0.2	-11.3	6.0	2.0
Hawaiian	-5.4	-1.1	13.8	-4.2	-3.2
Coronals					
Samoan	-4.3	0.0	6.5	-0.7	-1.5
Hawaiian	2.9	4	-2.7	-1.1	-3.1

While there is a marked difference between Hawaiian and Samoan with *a*, the vowel deviations on the whole are rather small and within the range found to be insignificant in Maddieson & Precoda (1992). This finding therefore is broadly consistent with Maddieson & Precoda's finding that consonantal place has little effect on the following vowel.

5. Vowel-vowel co-occurrence across syllables

V-V combinations across syllables have been studied statistically in Krupa (1966, 1967, 1971), which had the overall goal of the computing structural similarity between pairs of Polynesian languages and relating these similarity scores to genetic affiliation. In particular, Krupa examined the frequency of V-V pairs in disyllabic roots in eight Polynesian languages. Using an extension of the chi-square goodness of fit test, he found the statistically significant associations and disassociations of V1 and V2 shown below.

Table 9. Summary of Krupa 1971 V-V co-occurrence patterns

	i	e	a	o	u
i	+!	-			
e	-	+!			
a					
o				+	
u				-	+!

Statistically significant associations in four or more languages are indicated with a '+', significant associations in all eight languages with a '+!', and significant disassociations with '-'. The salient observation from Krupa's study is that non-low vowels tend to assimilate in height, producing a gradient pattern of total vowel assimilation. As these patterns are not found in alternations, we may alternatively think of them as static co-ccurrence patterns that disfavor two different non-low vowels with the same front/back specification. The sequence *o-u* is apparently excluded from this pattern (but see below). All of these patterns were found in Samoan, with the exception of *e-i*. The larger pattern here follows a common pattern in vowel harmony, either

categorical or gradient, in which a greater degree of similarity between target and trigger entails a greater degree of assimilation; see e.g., Hare (1990) on Hungarian vowel harmony.

The data organized below can be used to compare our data with Krupa's original study, as well as extending it to new empirical patterns. The V positions in the tables below may be filled either by singleton short vowels or a member of a diphthong that fits a V.(C)V profile, i.e., the second member in V1 position, or the first in V2 position. The heat maps below these tables map the O/E values to a hot-cold color scale, helping to visualize over- and under-representation in the lexicon.

Table 10. V.(C)V in disyllables

	i	e	a	o	u
i	52	5	26	30	12
e	13	33	30	10	14
a	52	57	122	73	55
o	28	26	41	70	16
u	33	24	61	5	55

Table 11. V.(C)V in all roots

	i	e	a	o	u
i	86	16	69	43	19
e	18	49	52	15	21
a	115	90	230	130	91
o	41	35	84	90	18
u	45	37	136	14	80

Table 12. V.(C)V O/E in disyllables

	i	e	a	o	u
i	2.20	0.26	0.70	1.20	0.60
e	0.69	2.15	1.01	0.50	0.87
a	0.77	1.03	1.14	1.02	0.95
o	0.82	0.93	0.76	1.94	0.55
u	0.98	0.88	1.15	0.14	1.92

Table 13. V.(C)V O/E in all roots

	i	e	a	o	u
i	1.97	0.49	0.84	1.03	0.58
e	0.62	2.26	0.95	0.54	0.96
a	0.93	0.98	1.00	1.10	0.98
o	0.81	0.93	0.89	1.87	0.48
u	0.77	0.85	1.24	0.25	1.82

Figure 3. Heat map of V.(C)V O/E disyll

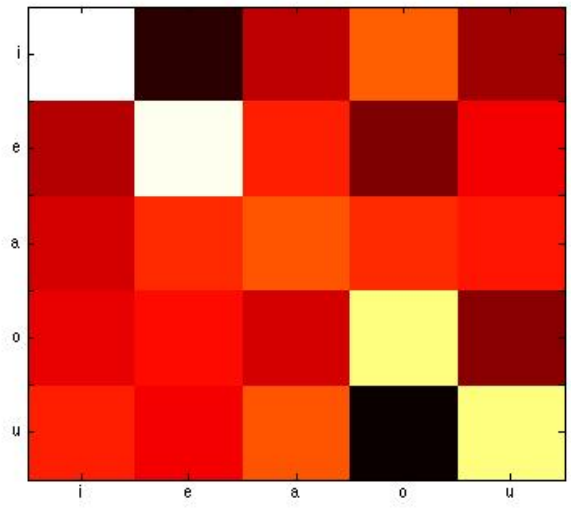
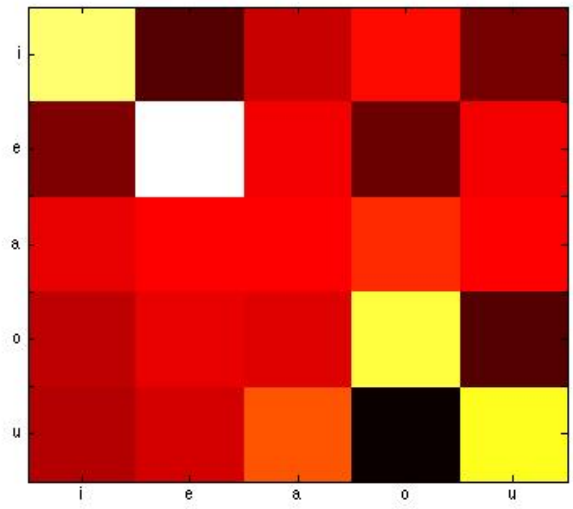


Figure 4. Heat map of V.(C)V O/E all roots



It is clear from this data that, like Krupa's study, there is a gradient pattern of vowel assimilation in which mid vowels are avoided when they combine with high vowels in the same front/back class. This is apparent from low counts and O/E values of non-identical V-V's in the boxed regions above. For comparison, these regions correspond to the same two-by-two boxes in Table 9, which summarizes Krupa's findings. Non-identical V-V pairs have comparatively low O/E values, while O/Es for identical vowels are rather high. However, all combinations with *a* approach 1.0, indicating a relative lack of deviation of O from E. These relationships can also be seen in the 'checker-board' patterns in the heat maps corresponding to the same boxes in the contingency tables. A chi-square goodness of fit test for the data in Table 10 (with disyllabic roots), the dataset most comparable to Krupa's (1967) Samoan dataset, shows that the null hypothesis that vowels occur independently can be rejected (χ^2=179.84, which is significant at a level higher than 0.001 with 16 degrees of freedom; cf. χ^2=103.22, reported in Krupa (1967)).

This dataset also suggests that Krupa's extended chi-square tests for goodness of fit may have missed certain patterns, as other combinations of mid vowels seem to be under-represented, and yet they are not analyzed as significant disassociations in Krupa's study. For example, while not as strong as *u-o*, *o-u* has O/E very low values in both disyllabic and all roots, i.e., 0.55 and 0.48 respectively, when compared with other V-V pairs. The same can be said of *e-i*, which has an O/E of 0.69/0.62. Also, one can observe a similar checker-board pattern with combinations of front + back vowels, as illustrated in the upper right regions of the two heat maps. While these associations and disassociations are clearly not as strong as the boxed V-Vs in Table 9, all of them have O/E's below 0.6 and they echo the patterns of parallel pairs with the same front/back specifications. However, the opposite order is curiously absent in non-low back + front vowels on the bottom left of the chart, indicating a role for vowel order.

Are these specific V-V combinations significant under- or over-representations? We have performed the same extension of the chi-square goodness of fit test for specific combinations employed in Krupa (1966, 1967), and found that the same significant patterns hold of the eight associations (*e-e, i-i, o-o, u-u*) and disassociations (*e-o, i-e, i-u, u-o*), but that other weaker patterns, i.e., *o-u, e-o, i-u*, have adjusted χ^2 values just below the significance threshold used by Krupa. While these latter patterns are not significant dissassociations in Krupa's tests, they might not be expected to be significant in other analyses. For example, it seems that the more similar two vowels are, the greater the under-representation of the pair: non-low vowels that agree on [+/-back], e.g., *i-e*, have lower O/E's than mid vowels from different [back] classes, e.g., *e-o.* In other words, while one can posit an absolute threshold in establishing the significance of a gap, e.g., an O/E of $n<0.7$, an alternative is to look for a correlation between O/E values and some other variable. The obvious choice is the continuous variable of phonological similarity (Pierrehumbert 1993, Frisch et al. 2004), because O/E appears to go down with greater similarity. While the number of data points is too small to document this pattern, such an approach can account for the fact that the more similar two vowels are, the more prone they are to this restriction. Another factor clearly apparent here is the order of vowels, e.g., *i-u* is under-represented when compared with *u-i.*

In summary, the quantitative study of Samoan roots has found the same pattern of non-low gradient vowel assimilation documented first in Krupa (1967), and it supports the extension of this pattern to vowel combinations logically implied by it, including *o-u*, *e-i*, and combinations of front + back vowels.

6. Consonant-consonant co-occurrence across syllables

Krupa's work on Polynesian languages has also clarified many interesting patterns for consonant combinations across syllables. Krupa (1971) provides a nice summary of this work, which examined combinations of consonants (and onsetless syllables) in disyllabic roots of eight Polynesian languages. The basic finding was that, for three different place of articulation classes, labial, coronal, and 'back' (=velar and glottal), most languages have significant tendencies to avoid consonant pairs in which both members of the pair are either coronal or labial. The lack of a homorganic restriction for so-called back consonants is claimed to be an innovation of Polynesian, as disassociations of this place class are found in Fijian and Proto-Austronesian (Krupa 1966). This last empirical claim is somewhat problematic, however, because it depends on certain assumptions about identical consonants and also incorrect claims about consonant classification, which are matters addressed below.

This investigation is of importance to many studies in generative phonology that attempt to formalize restrictions against homorganic consonant pairs. These restrictions have been found in several genetically unrelated languages, e.g., Arabic ((Greenberg 1950), (McCarthy 1988)), Russian (Padgett 1995), and Javanese ((Uhlenbeck 1950), (Mester 1986)). Indeed, Pozdniakov & Segerer (2007) argue that such restrictions are statistical universals of language. The question of how to formalize these restrictions is a matter of recent debate (Pierrehumber 1993, Frisch et al. 2004, Coetzee & Pater 2008, see Alderete & Frisch 2007 for a review). Our goal is not to weigh in on this debate, but rather to vigorously explore the co-occurrence data and organize the empirical patterns so that subsequent theoretical analysis can stand on stronger empirical ground. In particular, we examine a number of factors found to be important in other languages that were not examined in Krupa's study of Samoan roots. These factors include the distinction between identical and non-identical consonants, manner, and order of consonant occurrence.[2]

The two main root lists were used to quantify consonant co-occurrence, that is, the list of all roots, and the list of disyllabic roots, which has more overlap with Krupa's list. Frequencies of all possible consonant combinations in adjacent syllables were compiled and used to give the relative frequencies below, i.e., the O/E values for all combinations (see the data supplement for the actual frequencies). Onsetless syllables were also examined (see section 4 above), but they are not reported on here. The two place classes that have more than one member, labial and coronal, are boxed in these tables, and identical consonants are shaded to highlight the relative under-representation of non-identical homorganic consonant pairs. Next to each table is the corresponding heat map that helps visualize the relative frequency of a combination, where again black is the lowest point on the scale, white the highest. In the heat maps, the lighter boxes along the diagonal show the greater relative frequency of identical consonant pairs, and the dark regions of the labial-labial combinations show how restricted these combinations are.

[2] A careful reader of Krupa's research reports will note that manner and order effects are investigated by Krupa (his 'modal' and 'position' tests). However, they were not used as conditioning factors in the co-occurrence of homorganic consonants, which has been found to be important in several languages.

Table 14. O/E of adjacent C-C's in disyllables

	p	f	v	m	t	s	n	l	ŋ	ʔ
p	1.76	0.12	0.27	0.00	1.66	0.87	1.00	1.09	0.85	1.02
f	0.00	1.03	0.38	0.00	1.57	1.08	1.54	1.16	1.20	0.68
v	0.00	0.00	4.30	0.00	0.44	1.05	1.45	1.61	0.34	1.15
m	0.60	0.58	0.00	1.58	1.52	1.25	1.04	1.08	0.91	0.74
t	1.02	1.97	0.59	1.08	0.71	0.40	0.94	1.00	1.25	1.06
s	1.12	0.46	0.68	1.17	0.09	1.80	0.58	1.14	1.08	1.60
n	0.44	2.12	0.00	3.21	0.73	0.38	2.86	0.00	0.00	1.67
l	2.07	1.72	1.58	1.95	0.65	0.51	0.21	0.78	1.62	0.77
ŋ	0.45	0.88	0.00	0.00	1.25	2.76	0.33	0.84	2.68	0.43
ʔ	0.96	0.80	3.25	1.01	0.99	1.08	1.19	0.94	0.12	1.05

Table 15. O/E of adjacent C-C's in all words

	p	f	v	m	t	s	n	l	ŋ	ʔ
p	1.79	0.24	0.34	0.00	1.51	0.71	0.91	1.26	0.91	1.02
f	0.00	1.33	0.23	0.00	1.78	1.20	1.48	0.94	1.42	0.52
v	0.00	0.00	6.05	0.00	0.31	0.71	1.19	1.56	0.39	1.15
m	0.52	0.77	0.27	1.56	1.52	1.38	1.35	1.04	0.61	0.64
t	1.23	1.47	0.66	1.13	0.78	0.26	0.92	1.00	1.14	1.24
s	1.25	0.58	1.01	1.63	0.06	1.81	0.68	1.08	1.01	1.32
n	0.49	2.08	1.73	2.23	0.80	0.20	2.34	0.00	0.50	1.53
l	1.42	1.73	0.90	1.33	0.61	0.60	0.24	0.88	1.94	1.00
ŋ	0.85	0.73	0.00	0.00	1.09	2.86	1.02	0.66	2.13	0.51
ʔ	1.16	0.77	2.07	1.17	1.03	1.28	0.81	1.09	0.25	1.05

Figure 5. Heat map of C-C's in disyllables

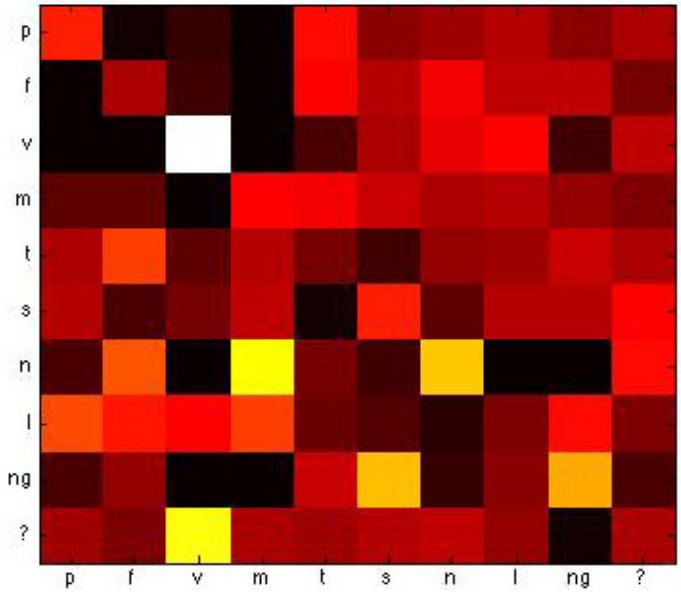
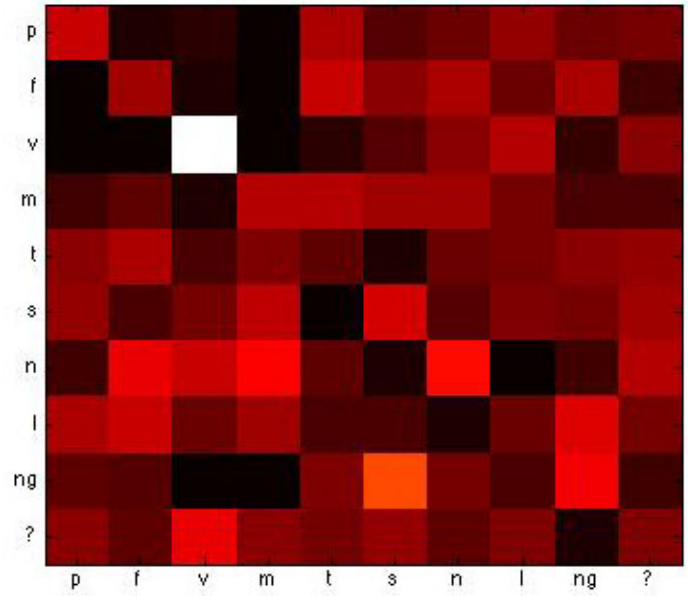


Figure 6. Heat map of C-C's in all roots



Let's begin with some basic observations that will lead to more specific points developed below. One general pattern is the relative under-representation of labial-labial and coronal-coronal combinations. For example, in the two heat maps, the four-by-four grid in the upper left regions corresponding to labial-labial combinations is far darker than other non-homorganic regions, illustrating the fact that they tend to have lower O/E values. Another important fact is that segment order matters, and it is necessary to examine specific consonant orders in describing consonant co-occurrence. For example, *m-p* and *m-f* have O/E's of 0.52 and 0.77 in all roots, respectively, but the opposite orders, *p-m* and *f-m*, both have 0.0, with no occurrences in both datasets. As for coronals, it seems that there is an effect of manner, because pairs of coronal obstruents and pairs of coronal sonorants have much lower O/E's than pairs with members from the two manner classes. This is evident in the heat maps from the checker-board pattern seen in the two-by-two grids for *t-s* and *n-l*. *s-n* pairs seem to be a kind of middle ground, despite being in different manner classes.

In comparing the O/E patterns in both lists (disyllabic and all roots), it appears that there are only very minor differences. The main differences surface in C-C pairs that have very high O/E's, like the difference between *v-v* in disyllables (4.30) and all roots (6.05); see also *n-m* and *ʔ-v*, which again have very high O/E's. The only exception to this statement is the contrast between *n-v* in disyllables, which is 0.0, and its high relative frequency in all roots, 1.73, a fact for which we have no explanation. Aside for these disparities, the differences between the two datasets are very slight, and so, unless otherwise noted, we use the larger root list in the rest of this section because it has a larger number of examples.

Returning to the issue of order, for the most part, the order facts seems to derive from the independent fact that bilabials are preferred in initial syllables while coronal sonorants are preferred in non-initial syllables (see section 4). Consider again the labial-labial combinations in Table 14, which are disyllabic and therefore the first consonant is always initial. All non-zero O/E values, with the exception of *f-v* (=0.23), involve either *m* or *p* in the first syllable. It is still a mystery why *m-v* is 0.0, but *v* is lowest frequency consonant (section 4), so it may just be a random gap. These numbers accord well with the qualitative statements about exceptions to a general prohibition on two labials in Mosel & Hovdhaugen (1992: 24-25). In this passage, *m-f, m-p, p-f,* and *p-v* are specifically singled out as occurring with a much higher frequency than other labial-labial sequences.

An alternative analysis for *m* initial C-C's is that our dataset has mistakenly included morphologically complex words, in particular roots with the prefix *ma-*. We have re-examined all of the *m*-initial C-C's, shown below, with this in mind. While it is a striking fact that all of them begin with the sequence *ma*, only one example*, maɟuta*, could plausibly have such a morphological analysis, where the *ma* might mark a comitative. Perhaps these forms derive historically from complex forms that contain a *ma-* prefix, but it would seem that the initial-syllable preference is a better explanation of the skewed distribution for labial-labial combinations because it is more general.

(2) Words exemplifying *m-p* and *m-f* pairs

a. Roots with *m-p*: *mapo* 'name of small fish (n)', *mapu* 'marble (n)', *mapeva* 'be sprained (v)', *mapo* 'be dry/firm (v)', *mapu* 'rest, have a break (v)', *mapu* 'whistle (v)', *mapuna* 'be strained'

b. Roots with *m-f*: *mafa* ‘pass, brow of a hill’, *mafu* ‘grated taro baked and cut into small pieces (n pl)’, *mafu* ‘fat of the breast of a pidgeon (n)’, *mafua* ‘tempting food … likely to draw game or fish (n)’, *mafia* ‘be think (v)’, *ma:fini* ‘grin, smile’, *mafu* ‘heal, dry up’, *ma:fua* ‘originate from …’, *ma:futa* ‘dwell, stay with s.o. (v)’, *mamafa* ‘be heavy’, *ma:ofa* ‘be amazed, astonish (v)’

Further support for this conclusion comes from the fact that coronal-coronal combinations likewise show higher O/E values when the two coronals differ in sonorancy and the coronal sonorant is in the preferred non-initial position. Limiting ourselves to disyllabic roots, to ensure that the first consonant is in the initial syllable, all of the pairs *n-t, n-s, l-t, l-s* have lower O/E’s than their corresponding pairs with the opposite orders where the coronal sonorant appears in the preferred non-initial position, e.g., *l-s* is 0.51, cf. *s-l*, 1.14. Thus, as with the ordering of bilabials, order matters for coronal sonorants.

Moving next to nasals, Mosel & Hovdhaugen (1992: 24) note in passing that the sequences *ŋVmV* and *ŋVnV* are systematically avoided, stating that combinations of nasals with *ŋ* seem to have low frequency. The contingency table below investigates these observations by giving the O/E’s of all nasal-nasal combinations in all roots. *ŋVmV* roots are clearly dispreferred, as are roots that end in *ŋ*. The relatively high frequency of *ŋVnV*, with an O/E of 1.02, may be an anomaly of the larger dataset because it is much lower in disyllabic roots: 0.33.

Table 16. O/E of adjacent nasal C-C’s

	m	n	ŋ
m	1.56	1.35	.61
n	2.23	2.34	.50
ŋ	0	1.02	2.13

A brief note on velars is relevant to restrictions on homorganic consonants. The combinations *ŋ-ʔ* and *ʔ-ŋ* have unusually low O/E’s, .51 and .25, respectively, in all roots, a pattern that is even stronger in disyllabic roots. It is quite likely that this is due to the fact that they were homorganic in Proto-Polynesian, as **k > ʔ* in Samoan. Furthermore, while Krupa (1966) does not report this in Maori for *ŋ-k* (*k* is retained in Maori), a reanalysis of the contingency table he provides of in Krupa (1968) also reveals a relatively low frequency of *ŋ-k*, with an O/E of 0.39. It seems, therefore, that combinations of velars are relatively under-represented in at least two Polynesian languages. This finding calls into question the conclusion in Krupa (1966) that Polynesian languages failed to continue a prohibition against homorganic consonants in all place classes.[3]

Finally, we return to the principal place classes, labials and coronals, to summarize some of the larger findings. The effects of order and manner (i.e., sonorancy) are shown below for the summed occurrences of each place class (we use actual occurrences here because the numbers are small). Orders 1 and 2 refer to the upper and lower triangles in Tables 14-15 (excluding identical consonants on the diagonal). There is a clear effect overall of manner: the counts of same-manner homorganic combinations are so low in comparison to the combinations in all

[3] Krupa’s (1967) conclusion that back consonants in Samoan, i.e., velars and glottals, do not have a co-occurrence restriction is further muddled by the inclusion of non-native /k/ in his root list.

roots. But this effect is not found in order 1 for labials. This is because all occurrences in order 1 involve *m* and there are no other labial sonorants.

Table 17. Frequencies of homorganic C-C's in Table 15, sorted by order

	Order 1		Order 2		Totals	
	lab	*cor*	*lab*	*cor*	*lab*	*cor*
All roots	6	129	21	30	27	159
Same sonorancy	6	5	0	5	6	10

It is difficult to extricate these effects of manner and order from the potential effect of place. For example, from the chart below, which scales C-C pairs by O/E values (top is most restricted, bottom, least restricted), it would appear that the restriction against homorganic C-C's is stronger for labials than coronals. For example, six of the twelve possible (non-identical) combinations of labials have a 0.0 O/E, while only two coronal combinations approach 0.0. On the opposite end of the scale, most different-manner coronal combinations, excluding *s-n*, approach an O/E of 1.0, the numerical value of a theoretically unrestricted C-C. Different-manner labials, on the other hand, range from 0.0 to .77, but do not continue to 1.0 as observed for coronals.

Table 18. O/E scale for all consonant pairs, sorted by place class and order

	Labial		Coronal	
	Order 1	**Order 2**	**Order 1**	**Order 2**
0	p-m, f-m, v-m [0]	f-p, v-p, v-f [0]	n-l [0]	
				s-t [.06]
.1				
.2	f-v [.23], p-f [.24]			n-s [.20]
		m-v [.27]	t-s [.26]	l-n [.24]
.3	p.v [.34]			
.4				
.5		m-p [.52]		
.6				l-s [.60], l-t [.61]
			s-n [.68]	
.7		m-f [.77]		
.8				n-t [.80]
.9			t-n [.92]	
> 1			t-l [1.0]	
			s-l [1.08]	

However, this initial impression may be spurious, because all comparisons of similar C-C pairs across place classes either also differ in order or they seem rather comparable. Thus, while it is true that *p-m* (0.0) and *m-p* (.52) have much lower O/E's than *t-n* (.92) and *n-t* (.80), these cases differ in that, with the first pair, both orders put a bilabial in the shunned non-initial syllable, but for *t-n* the coronal sonorant appears in the preferred non-initial position. Likewise, for *f-m* (0.0)

and *m-f* (.77), cf., *s-n* (.68) and *n-s* (.20); preferred orders could account for the relatively low or high numbers. Moreover, the only set of pairs that seem to factor out order, *p-f* (.24) and *f-p* (0.0) vs. *t-s* (.26) and *s-t* (.06), because they do not contain bilabials or coronal sonorants, seem rather comparable. In sum, the reduced inventory of Samoan just does not provide enough data points for one to document a clear difference between labials and coronals on the degree of non-occurrence.

7. Discussion

We have organized a list of 1137 disyllabic roots and a more comprehensive list of 1640 roots and explored the frequency structure of individual segments and combinations of segments. In particular, the frequency of syllable types, consonants, short/long vowels, and diphthongs are documented here. We have made certain observations about the the relative frequency of individual types, e.g., that the most frequent vowel is *a*, as well as observations about certain positional biases for segments. The positional bias found specifically for bilabials in initial syllables and non-labial sonorants in non-initial syllables provided additional precision to prior findings in Polynesian phonotactics. Frequencies of CV syllables were also examined and characterized by a lack of strikingly high or low frequency syllables, consistent with prior work on CV frequencies (Maddieson & Precoda 1992).

We also investigated a pattern of gradient vowel assimilation documented first in Krupa (1966, 1967, 1971) in which pairs of non-identical front or non-identical back vowels, excluding *a*, tend to be under-represented relative to corresponding pairs of identical front or back vowels. Our findings show that the patterns found to be significant in Krupa's research are also significant in a larger dataset, and further support the extension of the basic pattern to other perhaps weaker V-V restrictions, including additional orders and even gradient assimilation of pairs of front + back non-low vowels.

Finally, we investigated C-C pairs in adjacent syllables in some detail and gave more structure to the nature and scope of a restriction against homorganic C-C pairs, also previously investigated by Krupa. Once pairs of identical consonants were excluded, we found an effect of manner such that the homorganic restriction is stronger for C-C pairs in the same sonorancy class. Furthermore, we found that the specific order of consonants was crucial and related to the positional biases found in the analysis of segment frequencies. Generally speaking, the restriction against two same-place consonants was reduced when a bilabial was in the initial syllable and a coronal sonorant was in a non-initial syllable.

This report is intended primarily as a contribution to the description of Samoan phonotactics by giving a quantitative analysis of segment and segment combination frequencies. The results, however, have broader implications for the correct analysis of phonotactics in generative phonology. One important theme has been the role of similarity in the characterization of V-V and C-C restrictions. In both cases, it seems that, identical segment pairs aside, the more similar two segments are, the stronger the restriction or under-representation of the pair. Thus, pairs of non-low vowels that agree in front/back specification are more restricted than pairs that do not. In C-C pairs as well, consonant pairs that agree in sonorancy are far less common than those that do not.

This effect of similarity could be accounted for in a host of ways. For example, the effect of manner can be accounted for by positing constraints that specifically ban two consonants that

agree in place, as well as the feature [sonorant] (Suzuki 1998, Coetzee & Pater 2008), and weight these constraints accordingly. Alternatively, the effect of similarity could be derived from more general assumptions, like the conjecture formalized in Frisch (1996) and Frisch et al. (2004), that lexicons are shaped by lower level psycholinguistic processes that have difficulty processing sequences of similiar sounds. In Frisch (1996), this conjecture is based on the observation that the relative frequencies of consonant pairs correlates strongly with their phonological similarity, which is formalized as a ratio of shared phonological classes to the sum of shared and unshared natural classes (see also Alderete et al. (to appear, in press), for a formalization of similarity avoidance in a connectionist network).

Another important problem documented here is the impact of order in both V-V and C-C combinations. In V-Vs, certain orders have lower relative frequency than others. And in C-C's, the positional bias for bilabials and coronal sonorants is an important factor in the frequency structure of homorganic consonant pairs. How could order be incorporated in an analysis that attempts to formalize native speakers' intuitions about these restrictions? In the weighted constraints approach of Coetzee & Pater (2008), positional biases could simply be recognized as evidence for an independent set of constraints, e.g., *Bilabial/Non-InitialSyllable, *Coronal/InitialSyllable, etc., which effectively create a bias by directly banning the under-represented position for a segment class. The weight of each constraint, or its relative importance in the larger constraint system, could easily be learned with the standard error-driven learning algorithms that adjust weights in response to data. Thus, at least one tractable approach to factoring in order seems to exist, but the larger point to be made is that the correct analysis of the restrictions on combinations of segments really must factor in order somehow. With some exceptions, e.g., MacEachern (1999), the role of order is largely ignored in the analysis of consonant co-occurrence restrictions, and Samoan shows quite explicitly how important this is.

References

Alderete, John & Stefan Frisch. 2007. Dissimilation in grammar and the lexicon. The Cambridge handbook of phonology, ed. by P. de Lacy, 379-98. Cambridge: Cambridge University Press.

Alderete, John, Paul Tupper & Stefan Frisch. 2012a. Phonotactic learning without a priori constraints: Arabic root cooccurrence restrictions revisited. In press: Proceedings of the 48th annual meeting of the Chicago Linguistics Society. Chicago: University of Chicago.

Alderete, John, Paul Tupper & Stefan A. Frisch. 2012b. A connectionist approach to phonological constraint induction: OCP[Place] in Arabic. To appear: Language Sciences.

Berg, Rene van den. 1989. A grammar of the Muna language. Dordrecht: Foris.

Blevins, Julliette. 1995. Syllable in phonological theory. The handbook of phonological theory, ed. by J. Goldsmith, 206-44. Cambridge, MA: Blackwell.

Blust, Robert. 2007. Disyllabic attractors and the anti-antigemination in Austronesian sound change. Phonology.1-36.

—. 2009. The Austronesian languages. Canberra: Pacific Linguistics.

Coetzee, Andries & Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-ccurrence in Muna and Arabic. Natural Language and Linguistic Theory 84.289-337.

Elbert, Samuel H. 1953. Internal relationships of Polynesian languages and dialects. Southwestern Journal of Anthropology 9.147-73.

Francois, Alexandre. 2010. Phonotactics and the prestopped velar lateral of Hiw: resolving the ambiguity of complex segments. Phonology 27.393-434.

Frisch, Stefan. 1996. Similarity and frequency in phonology: Northwestern University Doctoral dissertation.
Frisch, Stefan A., Janet Pierrehumbert & Michael B. Broe. 2004. Similarity avoidance and the OCP. Natural Language and Linguistic Theory 22.179-228.
Greenberg, Joseph. 1950. The patterning of root morphemes in Semitic. Word 6.162-81.
Hare, Mary. 1990. The role of similarity in Hungarian vowel harmony: A connectionist account. Connectionist natural language processing, ed. by N. Sharkey, 295-322. Oxford: Intellect.
Harlow, Ray. 1991. Consonant dissimilation in Maori. Currents in Pacific Linguistics: Papers on Austronesian languages and ethnolinguistics in honour of George W. Grace. Pacific Linguistics Series C, no. 117, ed. by R. Blust, 117-28. Canberra: Australian National University.
Jones, ʻŌiwi Parker. 2008. Phonotactic probability and the Māori passive: A computational approach. Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2008), ed. by J. Eisner & J. Heinz, 39-48.
Krupa, Victor. 1966. The phonemic structure of bi-vocalic morphemic forms in Oceanic languages. The Journal of the Polynesian Society 75.458-97.
—. 1967. On phonemic structure of morpheme in Samoan and Tongan. Beitrèage zur Linguistik und Informationsverarbeitung 12.72-83.
—. 1968. The Maori Language. Moscow: Nauka Publishing House.
—. 1971. The phonotactic structure of the morph in Polynesian languages. Language 47.
—. 1973. Polynesian languages. The Hague: Mouton.
MacEachern, Margaret. 1999. Laryngeal cooccurrence restrictions. New York: Garland.
Maddieson, Ian & Kristin Precoda. 1992. Syllable structure and phonetic models. Phonology 9.45-60.
McCarthy, John J. 1988. Feature geometry and dependency: A review. Phonetica 43.84-108.
Mester, Ralf-Armin. 1986. Studies in tier structure: University of Massachusetts, Amherst Doctoral dissertation.
Milner, George Bertram. 1966. Samoan dictionary: Samoan-English, English-Samoan. London: Oxford University Press.
Mosel, Ulrike & Even Hovdhaugen. 1992. Samoan reference grammar. Oslo: Scandinavian University Press.
Padgett, Jaye. 1995. Stricture in feature geometry. Stanford, CA: CSLI Publications.
Pawley, Andrew K. 1966. Polynesian languages: a subgrouping based on shared innovations in morphology. The Journal of the Polynesian Society 75.39-64.
Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. NELS 23, 367-81.
Pozdniakov, Konstantin & Guillaume Segerer. 2007. Similar place avoidance: A statistical universial. Linguistic Typology 11.307-48.
Suzuki, Keiichiro. 1998. A typological investigation of dissimilation: University of Arizona Doctoral dissertation.
Uhlenbeck, E. M. 1950. The structure of the Javanese morpheme. Lingua 2.239-70.
Zuraw, Kie. 2000. Patterned exceptions in phonology: UCLA Doctoral dissertation.