To appear in Language and Linguistics 17, 2016.

Gradient vowel harmony in Oceanic

John Alderete, Sara Finley
Simon Fraser University, Pacific Lutheran University*

Abstract. This article contributes to the understanding of gradient phonological patterns by investigating graded vowel co-occurrence in Oceanic languages. In particular, vowel co-occurrence patterns in disyllabic stems are investigated in four languages: Samoan, Tongan, Hawaiian, and Fijian, as well as reconstructed forms in Proto-Oceanic and Proto-Malayo-Polynesian. With some variation in degree, all languages exhibit an over-representation of identical vowel pairs (e.g., *i*-*i*), an under-representation of similar vowel pairs (*i*-*e*), and no special restrictions on dissimilar vowel pairs (e.g., *i*-*o*). These graded restrictions are also subject to order effects in all languages because the dissimilar > similar inequality in frequency is only found in certain orders. Our focus is on documenting the patterns supporting these generalizations so that future theoretical analysis will rest on strong empirical ground. In addition, we propose one such analysis using gradient constraints on parasitic vowel harmony.

1. Introduction

As a convenient idealization, vowel harmony is often characterized as the categorical requirement that vowels share the same value of a specified feature, such as [back] or [round]. This type of requirement is standardly treated as across-the-board spreading from stems to affixes, or some other mechanism for ensuring featural agreement. Neutral vowels, i.e., vowels that do not undergo spreading, may complicate the picture, but straightforward assumptions about feature specification in autosegmental phonology seem to handle most cases, accounting for why transparent vowels are inactive and opaque vowels are active.

While vowel harmony is traditionally characterized as a morpho-phonological process, i.e., one that primarily affects affixes and specific form classes, there are reasons to believe that vowel harmony can also be applied to a model of phonotactics. First, vowel harmony is a phonetically motivated process (Archangeli & Pulleyblank 1994), suggesting that vowel harmony may become 'phonologized' on par with other phonotactic processes. Second, the same constraint-based mechanisms that are used to characterize vowel harmony generally apply without reference to morphological structure. This means that vowel harmony can be analyzed both as a morpho-phonological and phonotactic process. Third, vowel harmony has been incorporated into probabilistic models of phonotactics (see e.g., Goldsmith and Riggle (2012)), showing that even long-distance phonological patterns can be analyzed using probability distributions over vowel co-occurrences. The analysis of vowel harmony using probabilistic constraints has exciting implications for the analysis of a wide range vowel harmony patterns, including patterns that only resemble vowel harmony.

Most vowel harmony systems exhibit non-systematic behavior that cannot be straightforwardly explained using categorical representations of features and feature spreading. Many languages with rich affix morphology have invariant affixes that do not participate in harmony, e.g., Hungarian (Vago 1980). And most vowel harmony systems have a set of disharmonic roots that are more than just random exceptions to harmony generalizations. In Turkish roots, for example, /i e a o u/ may all co-occur freely, violating [back] and [round] harmony fundamental to the system, but /y ø ɨ/ must obey this harmony (Clements & Sezer 1982). Vowel harmony systems may also have pockets of irregularity that seem to be

* We are grateful to Robert Blust, Ashley Farris-Trimble, and Paul Tupper for their comments on an earlier version of this article, two anonymous reviewers from Language and Linguistics, Kim Myrhold, who helped organize our data, and Holly Wilbee, who assisted with data visualization. We alone are responsible for any errors that remain.

influenced by lexical distributions, as shown for back-neutral vowel sequences in Hungarian (Hayes & Londe 2006), and vowel co-occurrence patterns may exhibit language particular patterns that buck the usual trends predicted by generative phonology (Archangeli et al. 2012a; Archangeli et al. 2012b); see Harrison (1999) for a host of other factors leading to non-systematic vowel harmony. The existence of these non-categorical patterns, within seemingly categorical vowel harmony systems, suggests a range of vowel harmony patterns: from highly categorical to highly gradient. With the advent of models that make use of probability and gradience, it is now possible to analyze patterns that may be too weak to be categorized as vowel harmony under traditional assumptions in the same manner as one would analyze vowel harmony patterns that fall on the more traditional categorical side of the spectrum.

This article approaches graded vowel co-occurrence in certain Oceanic languages from this perspective. Oceanic languages do not have rich affix morphology, and very few have morphophonemic vowel harmony. Many of them, however, have stem co-occurrence restrictions that suggest an analysis involving gradient vowel harmony (see e.g., Krupa (1971)). We exemplify the restrictions of interest in Table 1 with vowel co-occurrence data from Samoan (Alderete & Bradshaw 2013). Vowel combinations are standardized using Observed/Expected values, which is a common measure used to assess over- (greater than 1.0) and under-representation (less that 1.0) in the lexicon (Pierrehumbert 1993). O/E for all pairs containing the low vowel *a* approaches 1.0, suggesting that the distribution of these pairs is not restricted. Non-low vowels, however, exhibit a gradient co-occurrence pattern in which pairs of identical vowels are significantly over-represented, similar vowel pairings are under-represented, and dissimilar vowels are intermediate on this scale. For example, *i-i* and *e-e* have O/E values above 2, but the non-identical pairs *i-e* and *e-i* have O/Es of 0.28 and 0.68, respectively, a clear contrast. These patterns of over- and under-representation of V-V pairs give rise to the 'checker board' patterns depicted in Figure 1, which visualizes the grades of these vowel sequences. We see the same patterns of over- and under-representation in the co-occurrence of /o u/, as shown in the bottom 2-by-2 region in the righthand corner of Figure 1. However, when we combine front vowels with back vowels, we only find an attenuated checker board pattern with combinations of front and back vowels, as shown in the upper righthand corner of this figure.

Table 1. Observed/Expected in V-Vs in Samoan disyllabic stems

	i	e	a	o	u
i	2.18	0.28	0.69	1.19	0.59
e	0.68	2.28	1.00	0.50	0.86
a	0.78	0.93	1.16	1.04	0.97
o	0.81	0.99	0.75	1.92	0.54
u	0.97	0.93	1.14	0.14	1.90

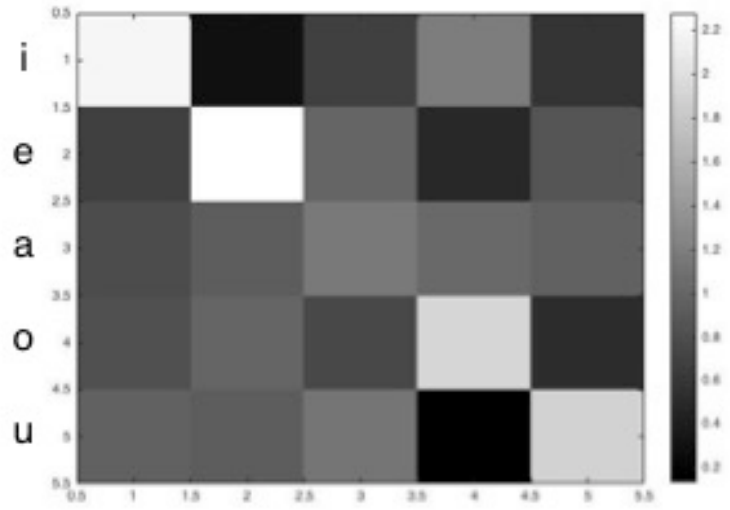

Figure 1. Heat map of Samoan V-Vs

While Samoan does not have active alternations producing these grades, it is possible to analyze the systematicity of the vowel co-occurrence facts as a product of a gradient form of vowel harmony. For example, the patterns of over- and under-representation above can be accounted for if the grammar of Samoan somehow formalizes a scale of acceptability such that identical vowel pairs are 'good', dissimilar pairs 'okay', and similar pairs are 'bad', e.g., *i-i* > *i-o* > *i-e*. One of the themes of this article is to show that these patterns resemble directional parasitic vowel harmony systems that exhibit both directionality effects and assimilation contingent on shared features (à la Cole and Trigo (1989)). Another theme is to illustrate the gradient nature of these patterns, and make the point that, like other types of non-systematic behavior (Frisch 1996), it is difficult to separate core (regular) patterns from exceptional (irregular) ones. We document gradient vowel harmony patterns like that illustrated above in four Oceanic languages, namely Samoan, Tongan, Hawaiian, and Fijian to establish an empirical basis for the gradient patterns. We also endeavor to examine the same patterns in reconstructed forms from Proto-Oceanic and Proto-Malayo-Polynesian, in order to support historical investigation of the potential source of these patterns. Our hope is that this historical and comparative perspective places our formal analysis on firm empirical ground.

The remainder of this article is organized as follows. In the next section, we investigate Samoan stem phonotactics further, and use this system to illustrate the methods employed for the phonotactic systems that follow. Section 3 provides some historical background for subsequent analysis, discussing the vowel systems of two ancestor languages and some phonological developments that are relevant to the analysis. Section 4 goes on to explore the stem phonotactics of three additional Oceanic languages, Tongan, Hawaiian, and Fijian. In section 5, we investigate vowel co-occurrence in Proto-Oceanic and Proto-Malayo-Polynesian, looking for the same phonotactic patterns to determine if the co-occurrence

patterns have been inherited. Section 6 gives a summary of the core facts across languages in order to sharpen the analytical focus. In section 7, we analyze the gradient patterns of parasitic harmony drawing upon constraints from Feature Spreading 2.0 (Jurgec 2011), formalized in Harmonic Grammar (Legendre et al. 1990; Pater 2009). In the conclusion, we sketch some of the ways that the core analysis may relate to harmony patterns found both within and outside the Austronesian language family.

2. Illustration of quantitative methods

This section explains the methods we use to extract generalizations about vowel co-occurrence by illustrating these methods for Samoan. In particular, we explain our methods for constructing stem lists, how contingency tables are constructed and analyzed from these lists, and the statistical tests used to establish the core generalizations.

The source for the Samoan data is Milner's (1966) lexicon. We extracted 1,871 content morphemes, i.e., nouns, verbs, and a small number of adjectives, from this source. Bound roots, adjectives, and words with diphthongs were excluded, producing a list of 1,512 nouns and verbs, of which 1,031 were disyllabic. We excluded adjectives and bound roots because we want a relatively homogenous set of word classes, and these exclusions did not significantly reduce the list. Further, diphthongs are excluded because we focus here on the distribution of simple vowels across syllables. The diphthongs of Samoan (*ei eu ai au ou oi ui*) complicate these simple vowel pairs because the two components of the diphthongs introduce locality issues that we wish to factor out. The focus on disyllabic stems is also prudent because prior research, e.g., Krupa (1971), used primarily disyllabic stems, so exclusion of stems that are not disyllabic allows us to compare results. Furthermore, stems are canonically disyllabic and stressed on the penultimate syllable in Oceanic, so longer stems often have unknown morphological complexity and introduce uncontrolled stress differences. Lastly, we note that Samoan is typical of Oceanic languages in having several homophones. We do include many homophones when they represent lexemes with a distinct etymology. However, a member of a set of homophones is excluded if it has only a slightly different meaning from other members of the set or differed only in part of speech, the idea being that these stems do not really constitute distinct lexemes, hence they are not distinct observations.[1]

The co-occurrence of vowels in disyllabic words (or "V-Vs" below) is given in the two sub-tables in Table 2. The rows in each table indicate the simple vowel of the first syllable, and columns indicate the vowel of the second syllable. Table 2a gives the raw frequencies, and Table 2b gives the Observed/Expected values, a standardized measure that allows us to compare the frequency of one cell relative to row and column totals. Following standard practice, O/E values that greatly exceed 1.0 are over-represented in the lexicon, and values that fall far below 1.0 are under-represented. Thus, the cells in Table 2b enable us to make observations like "*i*-*i* sequences is far greater than expected (2.18 > 1.0), compared to stems that contain *i* in the initial syllable, and stems that contain *i* in the second syllable".

[1] The full stem lists for Samoan and the other languages examined here are available from the first author's datasets webpage, which is: http://www.anderei.net/datasets/.

Table 2. Raw frequencies and O/E for Samoan disyllabic stems

2a.

	i	e	a	o	u
i	52	5	26	30	12
e	13	33	30	10	14
a	52	47	122	73	55
o	28	26	41	70	16
u	33	24	61	5	55

2b.

	i	e	a	o	u
i	2.18	0.28	0.69	1.19	0.59
e	0.68	2.28	1.00	0.50	0.86
a	0.78	0.93	1.16	1.04	0.97
o	0.81	0.99	0.75	1.92	0.54
u	0.97	0.93	1.14	0.14	1.90

In addition to O/E values, we also use inferential statistics to determine the reliability of the differences in frequencies. Following standard practice (Chomsky & Halle 1968), we form natural classes of vowels using three distinctive features: [back], [high], [low]. These features define three similarity classes of vowels of interest: identical (share three feature values), similar (share two features values), and dissimilar (share one feature value). Since we do not examine the low vowel *a* in any of these tests, we do not compare any vowels that have no features values in common at all. For the non-low vowels, two additional distinctions in natural class are relevant: front vs. back, and high vs. mid. In Table 3, we show different counts of identical versus similar vowels, aggregated by different feature classes. For example, there are 85 instances of identical front vowel pairs (52 *i-i* + 33 *e-e*) and 18 instances of similar front vowel pairs (13 *e-i* + 5 *i-e*) in the first row of Table 3a.

Two chi-square tests are reported in these tables: a one-way goodness of fit (GoF) test and a chi-square on the entire contingency table (CT).[2] Since the rows represent simple two-way contrasts, we can assume that, all else being equal, the two classes of vowel sequences should have a 50%-50% split.[3] The GoF test looks at each row and assesses how well this 50-50 statistical model accounts for the data. The GoF tests are significant in all rows in Table 3, indicating that the 50-50 split model can be rejected for all contrasts. For example, the 50-50 split model can be rejected as a good predictor of the distributions of identical front vowels and similar front vowels, which are 85 and 18 respectively. The rationale here is that these classes deviate significantly from the 50-50 split, requiring explanation. The CT test on the entire contingency table tests for an association between the categories indicated in the rows and columns. In Table 3a, for example, the CT test looks for an association between the front/back contrast and similarity (identical vs. similar). Is the contrast between identical and similar vowel sequences (V-Vs henceforth) only found in one front-back class, for example, or are there different contrasts for front vowels and back vowels? The CT test in Table 3a is not significant ($P = 0.6283$), so there is no association between tongue-advancement and similarity. Looking at the broader results, Tables 3a and 3b show strong differences in the similarity classes, but no association with the feature classes [back] or [high] in either CT test. Thus, identical front vowel pairs in Table 3a outnumber similar pairs by close to a 4-to-1 ratio in Table 3a, and approximately a 3-to-1 ratio in Table 3b, which supports the rejection of the hypothesis that the two similarity classes have a roughly equal distribution. However, this difference between identical and similar V-Vs is not associated with a particular [αback] class, because the CT test is not significant. The same is true of the similarity effect sorted by height classes reported in Table 3b.

[2] All tests were two-tailed and have one degree of freedom. The CT test used Yates correction.

[3] One might object to the assumption that the two classes should have a 50-50% split because a single low frequency vowel might on its own lead to rejecting the null hypothesis. We checked for this by conducting a CT test on the inputs to the GoF tests, e.g., [55, 5; 13, 33] for Samoan /i e/, and found that all of the results were the same, except two of 48 tests in the article were marginal rather than significant.

Table 3. Chi-square tests for effect of similarity, counts by feature classes

3a. Similarity, front/back			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	0.23	0.6283
Front	85 (*ii* + *ee*)	18 (*ie* + *ei*)	GoF	43.58	< 0.001 *
Back	125 (*uu* + *oo*)	21 (*uo* + *ou*)	GoF	74.08	< 0.001 *
Totals	210	39	GoF	117.43	< 0.001 *

3b. Similarity, high/mid			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	0.33	0.5662
High	107 (*ii* + *uu*)	45 (*iu* + *ui*)	GoF	25.29	< 0.001 *
Mid	103 (*ee* + *oo*)	36 (*eo* + *oe*)	GoF	32.30	< 0.001 *
Totals	210	81	GoF	57.19	< 0.001 *

We also examined the difference between pairs of similar and dissimilar vowels, i.e., vowels that share the value for two features (similar) or just one (dissimilar). These tests are motivated by the clear effect of order we find in vowel sequences: back-front V-Vs are much more common than front-back, and the similarity effect seems to be sensitive to the order of height classes. Table 4 applies the same statistical tests from above to similar-dissimilar classes. We find here clear associations with the two order classes. That is, the contrast between similar and dissimilar V-Vs is affected by order. Thus, similar and dissimilar pairs have a 1-to-2 ratio in front-back vowel pairs, i.e., *i-u* and *e-o* versus *i-o* and *e-u,* but the opposite order has a roughly equal distribution (Table 4a). Likewise, dissimilar pairs occur much more frequently than dissimilar pairs in high-mid sequences, but not so for mid-high sequences, again showing an association with the order of height classes (Table 4b).

Table 4. Chi-square tests for effect of similarity, counts by order classes

a. Similarity, front/back order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	5.78	0.0162 *
Front-Back	22 (*iu* + *eo*)	44 (*io* + *eu*)	GoF	7.33	0.0068 *
Back-Front	59 (*ui* + *oe*)	52 (*oi* + *ue*)	GoF	0.44	0.51
Totals	81	96	GoF	1.27	0.26

b. Similarity, high/mid order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	9.23	0.0024 *
High-Mid	10 (*ie* + *uo*)	54 (*io* + *ue*)	GoF	30.25	< 0.001 *
Mid-High	29 (*ei* + *ou*)	42 (*oi* + *eu*)	GoF	2.38	0.12
Totals	39	96	GoF	24.07	< 0.001 *

In sum, there are clear differences between the occurrence of identical and similar, on the one hand, and similar and dissimilar vowel pairs. However, the latter difference is only statistically significant in certain orders, namely front-back and high-mid orders.

These relationships and associations are visualized in Figure 2, which plots O/E over the similarity classes for specific vowel pairs. Pairs of identical vowels are clearly over-represented, occurring much more frequently than pairs of both similar and dissimilar vowels. Pairs of dissimilar vowels approach an O/E of 1.0, suggesting that they are neither over- or under-represented. With the exception of *u-i* and *o-e*, pairs of similar vowels are under-represented, with many falling drastically below 1.0. By comparing the specific V-Vs in the similar column, we can also see the order effect. Thus, for similar vowels only, there seems to be a preference for mid-high sequences; compare vowels on the left of the '2' SharedFeature column. Likewise, there is a preference for back-front V-Vs (right side). The magnitude of both of the order effects is about a 0.3-0.4 difference in O/E.

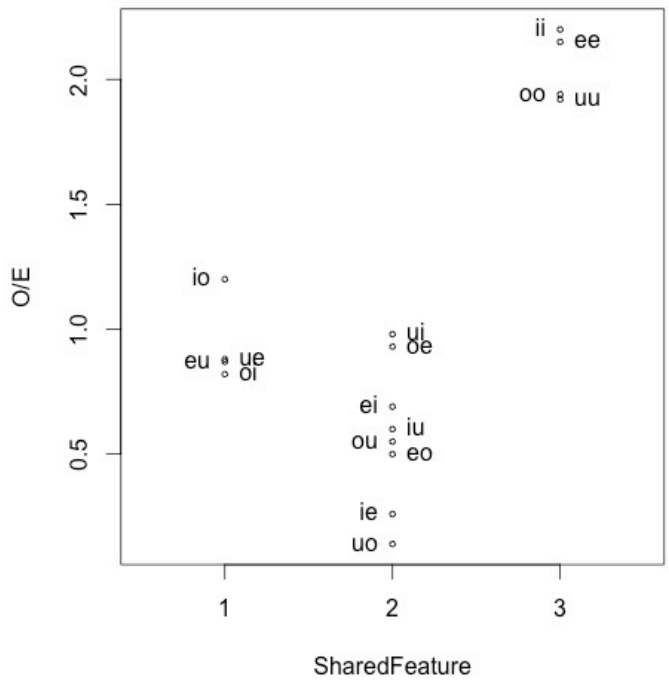

Figure 2. Similarity avoidance and order effects in Samoan

In sections 4 and 5, we examine data collected in essentially the same way and perform these same tests, both to provide additional support for these empirical generalizations and compare how they apply across languages.

3. Historical background

Below we provide some historical perspective on the vowel co-occurrence patterns, examining vowel developments from Proto-Malayo-Polynesian, vowel harmony in the larger Austronesian language family, and some conditioned phonological developments that are relevant to the central empirical claims.

Guided by prior research, Blust (2009/2013) reconstructs the following vowel systems for Proto-Malayo-Polynesian (PMP) and Proto-Oceanic (POC). The four synchronic vowel systems we investigate are within the Central Pacific subgrouping, which is in turn grouped within Oceanic. An awareness of the ancestral vowel systems of PMP and POC may therefore provide some perspective on these present day languages.

Table 5. PMP and POC reconstructed vowels

PMP	POC
*a	*a
*ə	*o
*i	*i
*u	*u
*-ay	*e
*-aw	*o
*-uy	*i
*-iw	*i

PMP *ə is the least stable vowel and has many reflexes in Malayo-Polynesian languages other than POC *o. In contrast, the vowels *a, *i, and *u tend to be rather stable and have been preserved in at least some environments in all languages. While 'diphthongs' are rather rare in our PMP dataset, they did lead to some mid vowels in Proto-Oceanic, as shown above. These developments raise the question of whether the under-represented V-Vs derive directly from the low frequency of vowels in an ancestor language. After all, all of the significantly under-represented V-Vs contain mid vowels, which are the reflexes of secondary vowels in PMP, i.e., *ə or diphthongs. It could just be that the V-Vs in question are under-represented because the secondary vowels were under-represented in POC and PMP. We investigate vowel co-occurrence in both POC and PMP in section five, but it is useful to quickly examine the raw frequencies of these vowels in these proto-languages to address this issue.

The frequencies of the five simple vowels of PMP, POC, and Samoan are given below (data sources are given in section 2 for Samoan, and sections 5.1 and 5.2 for POC and PMP respectively; see also Chrétien (1965), Table 3, for data from Proto-Austronesian that is broadly similar to the PMP data below). It turns out that this historical perspective is really only useful in contextualizing *e*, which clearly had a low frequency in POC (the reconstructed sources in PMP for *e* are too complex to represent here, but they would presumably show the same point). On the other hand, the frequency of the mid vowel *o*, and its main source from PMP, *ə, are comparable with the frequency with all other non-low vowels. It is even the second most frequent vowel in Samoan, which rules out a simple historical account of *o*.

Table 6. Vowel frequencies in PMP, POC, and Samoan disyllabic stems, sorted by stem position (1st/2nd syllable)

	i		**e**		**a**		**o**		**u**		
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	1st	2nd	
PMP	*20.02*		*(omitted)*		*32.97*		*(*ə) 19.45*		*27.56*		*% total*
	353	317			551	551	382	268	418	503	
POC	*16.85*		*4.48*		*34.31*		*19.84*		*24.52*		*% total*
	118	130	17	49	268	237	133	159	200	161	
Samoan	*16.24*		*12.59*		*33.71*		*19.77*		*17.68*		*% total*
	125	178	100	135	349	280	181	188	178	152	

The crucial point, however, is that this historical account of *e* does not diminish the importance of its over- and under-representation in the vowel co-occurrence patterns we examine. In our chi-square tests, the frequencies never dip below the threshold for a valid test, and, more to the point, the frequencies are standardized, and so they give relative measures of vowel co-occurrence. For example, Samoan *i-e* has a low Observed/Expected value of 0.27 (see data from section 2), because the actual

observed number of forms is low, relative to the frequencies of *CiCVC* and *CVCeC* stems. Likewise, while there are only 33 *CeCeC* stems, the O/E for *e-e* is quite high at 2.28 because this is far above the expected vowel for such stems. It seems clear in the Samoan case, therefore, and even in POP and PMP systems discussed below in section 5) that the behavior of both mid vowels call out for an analysis.

While none of the languages examined here has morphophonemic vowel harmony, such rules are attested, albeit rarely, in the language family (see Blust 2009/2013: 257 ff.). A quick survey of these rules and certain harmony-like phonotactic generalizations is therefore useful to our understanding of vowel co-occurrence in the family generally. A handful of languages exhibit prefixes that harmonize with the first vowels of the stem, including a few languages in which prefix vowels assimilate completely to the stem vowel, e.g., Balatank (eastern Celebic) *maŋa-wawau* 'to do', *meŋe-memeli* 'to cool', etc (Busentiz & Busenitz 1991). Total assimilation of this kind is also attested in Kapampangan (Philippine, central Luzon) and Seediq (Formosan, Atayalic), and similar patterns with a subset of prefix vowels are documented in Banggai (eastern Celebic). Again, while rare, progressive harmony from a prefix or particle may also lead to alternations in stems, like in Chamorro fronting harmony: *i gima?* 'the house', cf. *guma?*.

Many Austronesian languages also exhibit co-occurrence restrictions that avoid disharmonic vowel sequences, like our Oceanic case studies here. For example, Sabah languages avoid *oCa* sequences, usually favoring *aCa* (Kroeger 1992), and Northern Sarawak languages have a strong tendency to avoid [+high] + [-high] vowel sequences when the two vowels agree in [back] (Blust 2009/2013), a well-known type of height harmony. These alternations and static restrictions give perspective on our results here. For example, the restriction against *high-mid vowels in Northern Sarawak languages is particularly interesting because we find the exact same order effect in Oceanic languages; see section 2 and the discussion below.

Focusing now on vowel sequences, many of the co-occurrence patterns in stems we examine below have been extensively researched in a series of articles by V. Krupa (Krupa 1966; Krupa 1967; Krupa 1971); see also Chrétien (1965) for a rather cursory account of the same facts in Proto-Austronesian. Krupa's main goal was to examine over- and under-representation of specific V-V sequences and use them as features in a similarity metric for establishing conclusions about the classification of Oceanic languages. While we are not directly concerned with language classification, it is useful to review his findings and some of the limitations of them. Table 7 summarizes his results for eight Polynesian languages; '+' indicates an association between the two vowels in most of the languages, '-' a disassociation in most of the languages, and parentheticals mean that these effects were found in only two of the eight languages. From these patterns, it is clear that Krupa's larger findings are broadly similar to our sketch of Samoan above. There are no (dis)associations with *a*, and we find the same checker board pattern in pairs of similar front or back vowels (compare with Figure 1), and an echo of that pattern in front-back pairs.

Table 7. Summarization of Krupa's (1971: Table 17) co-occurence findings

	i	e	a	o	u
i	+	-		(+)	-
e	-	+		(-)	
a					
o				+	
u				-	+

Krupa's findings are an important precedent to this study in that they document many of the empirical patterns we grapple with here. However, there are some empirical and conceptual problems that prevent us from relying on Krupa's findings. As will become clear in our case studies below, Krupa's specific findings do not always line up with ours. For example, Krupa claims that Tongan bucks

the usual Polynesian trend of shunning *u-o* and *e-i*, and that any universal account must allow for this kind of language particular variation. However, Tongan does seem to have disassociations with these vowels in our dataset (see below), a fact that could either be due to different datasets or the specific methods used to analyze them. Another empirical discrepancy is that Krupa did not find a disassociation in *o-u* in any language, but we found them in all of our case studies. In general, Krupa's approach is to examine individual V-V combinations and look for an association/disassociation between specific vowels. Thus, every V-V sequence serves as a categorical feature for comparing languages. Our chi-square tests are categorical too. However, we examine sets of V-Vs here, and we also look for graded differences by examining O/E within an entire contingency table. This approach tends to reveal gradient differences between identical and similar, and similar and dissimilar V-V sequences that are not found in Krupa's approach. Finally, we examined order effects in front/back and high/mid classes that reveal important gaps in the similarity effect. In sum, we build on Krupa's pioneering work here by documenting vowel co-occurrence with more reliable data and methods that allow us to see new patterns.

Finally, we must comment on a few sound changes that bear directly on our empirical claims. The first concerns a sporadic change involving total assimilation that may have had an impact on identical V-Vs. Elbert (1953) created a cognate list of 202 words from 20 Polynesian languages and found many instances of total vowel assimilation in which a sequence of two non-identical vowels in Proto-Polynesian developed into a sequence of identical vowels, e.g., PPN **keli* > Tikopia *kere* 'dig'. In particular, Elbert found 33 cases of total assimilation in the 202 words in at least one Polynesian language. It is not possible from Elbert's dataset to calculate the percentage of total assimilations in any particular language, but these sound changes relate directly to our research focus because they may account for the over-representation of identical V-Vs. Scrutiny of the attested forms, however, suggests that this is unlikely. First, the percentage of total assimilations in any particular language is exceedingly low. Second, the input and output V-Vs do not seem to pattern with the vowel co-occurrence data we find in the daughter languages. For example, the "most marked" pattern, i.e., the input V-V that undergoes total assimilation the most often, is *a-o,* accounting for ten of the 33 cases, and 18 of the 33 cases, roughly half, contain *a* as either the first or second vowel. But V-Vs with *a* are not under-represented in any language. Furthermore, *a-a* is the most common output pattern of total assimilation, accounting for 11 of the 33 cases, cf. 10 outputs for *u-u*, 6 for *o-o*, 5 for *e-e*, 3 for *i-i*. Thus, we have many sporadic total assimilations with *a,* and comparatively fewer with non-low vowels, which is the opposite of what we expect from Krupa's work and our sketch of Samoan above. Clearly, additional factors are at work.

There also seem to be several widespread sound changes in Austronesian languages that actually reduce the frequency of identical V-Vs. One is low vowel dissimilation, a conditioned change in some Oceanic languages that raises *a* when another *a* occurs in the following syllable (Blust 1996; Lynch 2003). This change had the effect of reducing *a-a* in stems. Oceanic languages also had sporadic fronting of **u* to *i*, e.g., PPN **qumu* > Hawaiian *imu* 'earth oven' (Blust 1970), which reduced the overall frequency of *u-u* sequences. Finally, in a host of conditioned changes in Austronesian languages, *ə has been reflected as *i, u, o,* and *a* in penultimate and ultimate syllables (Blust 2009/2013), which effects stems that would have otherwise developed into *o-o*. We have not examined the impact of all of these in the languages we examine, but we believe that taken together these sound changes must have had the effect of reducing identical V-Vs in Oceanic. This conclusion makes the observation that they are over-represented in the lexicons of many languages all the more interesting.

4. Case studies: Synchronic systems

Below we explore the same vowel co-occurrence patterns in three additional Oceanic languages, Tongan (Polynesian, Tongic), Hawaiian (Nuclear Polynesian), and Fijian (East Fijian-Polynesian). These languages and Samoan belong to the Central Pacific sub-grouping within Oceanic. The languages were selected because they have reasonably good dictionaries and they represent discrete systems within Central Pacific.

4.1 Tongan

The methods for extracting Tongan stems are essentially the same as those used above for Samoan. An initial list of 2,190 content words was compiled from Churchward (1959), and this list was reduced to 1,964 after bound roots and words with diphthongs were removed. Analysis of disyllabic stems produced the vowel co-occurrence data in Table 8. Comparisons between similar and identical V-Vs, on the one hand, and similar and dissimilar V-Vs, are reported in Tables 9 and 10.

Table 8. Raw frequencies and O/E for Tongan disyllabic stems

	i	e	a	o	u	i	e	a	o	u
i	62	8	43	38	23	1.87	0.32	0.80	1.25	0.73
e	21	58	49	20	24	0.64	2.32	0.93	0.66	0.77
a	64	64	158	62	78	0.79	1.03	1.21	0.83	1.01
o	48	35	80	97	39	0.84	0.81	0.87	1.85	0.72
u	62	31	85	19	81	1.17	0.77	0.99	0.39	1.60

Table 9. Chi-square tests for effect of similarity, with feature classes

a. Similarity, front/back			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	1.09	0.2968
Front	120 (*ii* + *ee*)	29 (*ie* + *ei*)	GoF	55.58	< 0.001 *
Back	178 (*uu* + *oo*)	58 (*uo* + *ou*)	GoF	61.02	< 0.001 *
Totals	298	87	GoF	115.64	< 0.001 *

b. Similarity, high/mid			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	5.68	0.0171 *
High	143 (*ii* + *uu*)	85 (*iu* + *ui*)	GoF	14.75	0.0001 *
Mid	155 (*ee* + *oo*)	55 (*eo* + *oe*)	GoF	47.62	< 0.001 *
Totals	298	140	GoF	57.00	< 0.001 *

Table 10. Chi-square tests for effect of similarity, with order classes

a. Similarity, front/back order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	1.78	0.18
Front-Back	43 (*iu* + *eo*)	62 (*io* + *eu*)	GoF	3.44	0.064
Back-Front	97 (*ui* + *oe*)	98 (*oi* + *ue*)	GoF	0.01	0.942
Totals	140	160	GoF	1.333	0.282

b. Similarity, high/mid order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	6.358	0.0117 *
High-Mid	27 (*ie* + *uo*)	69 (*io* + *ue*)	GoF	18.38	< 0.001 *
Mid-High	60 (*ei* + *ou*)	72 (*oi* + *eu*)	GoF	1.09	0.30
Totals	87	141	GoF	12.78	< 0.001 *

In front and back vowel pairs in Table 9a, identical V-Vs greatly outnumber similar V-Vs by approximately 3- or 4-to-1, but this disparity is split in the height classes where the ratios are much higher with mid vowels than high vowels, hence the association with height class in Table 9b. The similar-dissimilar comparison aggregated by orders shown in Table 10 is also similar to Samoan, except the GoF test did not support rejection of the null hypothesis in front-back V-Vs. However, the patterns are in line with Samoan, with a significant result for high-mid V-Vs and the same trend as Samoan in front-back pairs. These results support the claim that there is a graded continuum among identical, dissimilar, and similar V-Vs, and the order effects are in the same direction, favoring back-front and mid-high orders.

4.2 Hawaiian

We extracted 12,046 nouns and verbs from (Pukui & Elbert 1986). After removing words with diphthongs, 9,493 words remained, 2,277 of which were disyllabic. The raw vowel co-occurrence patterns are given in Table 11, and Tables 12-13 compare identical and similar V-Vs, and similar and dissimilar V-Vs, respectively.

Table 11. Raw frequencies and O/E for Hawaiian disyllabic stems

	i	e	a	o	u	i	e	a	o	u
i	109	22	58	102	34	1.75	0.43	0.56	1.73	0.70
e	18	94	115	49	41	0.30	1.87	1.14	0.85	0.86
a	75	64	234	64	56	0.80	0.82	1.49	0.72	0.76
o	66	64	83	112	38	0.95	1.11	0.72	1.70	0.70
u	83	47	95	6	107	1.28	0.88	0.88	0.10	2.11

Table 12. Chi-square tests for effect of similarity, with feature classes

a. Similarity, front/back			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	0.01	0.94
Front	203 (*ii* + *ee*)	40 (*ie* + *ei*)	GoF	109.34	< 0.001 *
Back	219 (*uu* + *oo*)	44 (*uo* + *ou*)	GoF	116.44	< 0.001 *
Totals	422	84	GoF	225.78	< 0.001 *

b. Similarity, high/mid			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	0.01	0.0059
High	216 (*ii* + *uu*)	117 (*iu* + *ui*)	GoF	29.43	< 0.001 *
Mid	206 (*ee* + *oo*)	113 (*eo* + *oe*)	GoF	27.11	< 0.001 *
Totals	422	230	GoF	56.44	< 0.001 *

Table 13. Chi-square tests for effect of similarity, with order classes

a. Similarity, front/back order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	18.25	< 0.001 *
Front-Back	83 (*iu* + *eo*)	143 (*io* + *eu*)	GoF	15.93	< 0.001 *
Back-Front	147 (*ui* + *oe*)	113 (*oi* + *ue*)	GoF	4.45	0.035
Totals	230	256	GoF	1.39	0.24

b. Similarity, high/mid order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	15.67	< 0.001 *
High-Mid	28 (*ie* + *uo*)	149 (*io* + *ue*)	GoF	82.72	< 0.001 *
Mid-High	56 (*ei* + *ou*)	107 (*oi* + *eu*)	GoF	15.96	< 0.001 *
Totals	84	256	GoF	87.01	< 0.001 *

Like Samoan and Tongan, identical V-Vs greatly outnumber similar by about 5-to-1 when sorted by front/back classes (Table 12a), and 2-to-1 sorted by height classes (Table 12b). Dissimilar pairs outnumber similar ones only in front-back V-Vs; back-front V-Vs actually have the opposite inequality, where similar V-Vs outnumber dissimilar (Table 13a). However, dissimilar V-Vs greatly outnumber similar V-Vs across the board when ordered by height categories (Table 13b). Apparently, Hawaiian does not have as strong a preference for mid-high orders as other languages examined here.

4.3 Fijian

A list of 3,720 content words from Capell (1941/1957/1968) was reduced to 3,311 after exclusions. 2,142 are disyllabic, and their vowel co-occurence patterns are tabulated in Tables 14, 15, and 16.

Table 14. Raw frequencies and O/E for V-Vs in Fijian disyllabic stems

	i	e	a	o	u	i	e	a	o	u
i	114	12	96	55	12	1.95	0.27	1.04	1.19	0.25
e	37	118	75	14	37	0.65	2.71	0.83	0.31	0.81
a	130	71	249	86	90	1.03	0.73	1.24	0.86	0.88
o	72	75	122	147	39	0.78	1.07	0.84	2.02	0.53
u	60	40	112	25	155	0.76	0.66	0.89	0.40	2.43

Table 15. Chi-square tests for effect of similarity, with feature classes

a. Similarity, front/back			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	0.01	0.9203
Front	232 (*ii* + *ee*)	49 (*ie* + *ei*)	GoF	119.18	< 0.001 *
Back	302 (*uu* + *oo*)	64 (*uo* + *ou*)	GoF	146.33	< 0.001 *
Totals	534	113	GoF	273.94	< 0.001 *

b. Similarity, high/mid			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	1.36	0.2435
High	269 (*ii* + *uu*)	72 (*iu* + *ui*)	GoF	113.81	<0.001 *
Mid	265 (*ee* + *oo*)	89 (*eo* + *oe*)	GoF	87.5	<0.001 *
Totals	534	161	GoF	200.19	<0.001 *

Table 16. Chi-square tests for effect of similarity, with order classes

a. Similarity, front/back order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	33.1	<0.001 *
Front-Back	26 (*iu* + *eo*)	92 (*io* + *eu*)	GoF	36.92	<0.001 *
Back-Front	135 (*ui* +*oe*)	112 (*oi* + *ue*)	GoF	2.14	0.1433
Totals	161	204	GoF	5.07	0.0244 *

b. Similarity, high/mid order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	5.16	0.0231
High-Mid	37 (*ie* + *uo*)	95 (*io* + *ue*)	GoF	25.49	< 0.001 *
Mid-High	76 (*ei* + *ou*)	109 (*oi* + *eu*)	GoF	5.89	0.0153 *
Totals	113	204	GoF	26.12	< 0.001 *

As expected, identical V-Vs outnumber similar V-Vs by about 5-to-1 when sorted by front/back classes and 3- or 4-to-1 when sorted by height. Dissimilar V-Vs also greatly outnumber similar V-Vs by approximately 3-to-1, though only significantly in the front-back and high-mid orders, as we have seen in the three other languages. Fijian is like Hawaiian in reversing the usual dissimilar > similar trend by having a higher number of similar V-Vs than dissimilar V-Vs in the order back-front.

5. Case studies: Proto-languages

Are these patterns also observed in the ancestor languages that developed into these languages? We explore the same patterns in Proto-Oceanic and Proto-Malayo-Polynesian below. Our investigation here supplements the vowel co-occurrence patterns documented in Chrétien (1965) for Proto-Austronesian.

5.1 Proto-Oceanic

The source for our reconstructed Proto-Oceanic (POC) forms is the Austronesian Comparative Dictionary (Blust & Trussel 2013; Blust & Trussel 2015 (ongoing)), the most comprehensive and well-described set of linguistic reconstructions for Austronesian languages. 1,030 reconstructed proto-forms were extracted form this website (http://www.trussel2.com/acd/) and processed using a procedure similar to the one described in section 2, with some modifications. For variant forms, we simply took the first variant if the vowels are the same as other variants. The Austronesian Comparative Dictionary (ACD) contains many morphologically complex words, and many of them, e.g., causatives or frequentatives, have an obvious morphological relationship with another word in the dataset. We scrutinized the wordlist to find such relationships and only took the stem of one member of a word family. After these exclusions, and removing words with diphthongs, the initial list was reduced to 832 words, 736 of which were disyllabic. Tables 17, 18, and 19 show the vowel co-occurrence data for these words. The vowel types in POC are rather stable and essentially the same as the types in the daughter languages, as explained in section 3, so we can compare them directly with the patterns in section 4. However, with the smaller sample, we cannot have as much confidence in these findings.

Table 17. Raw frequencies and O/E for V-Vs in Proto-Oceanic disyllabic stems

	i	e	a	o	u	i	e	a	o	u
i	36	3	34	26	19	1.73	0.38	0.89	1.02	0.74
e	0	8	6	1	2	0.00	7.07	1.10	0.27	0.54
a	51	23	95	55	44	1.08	1.29	1.10	0.95	0.75
o	11	6	37	63	16	0.47	0.68	0.86	2.19	0.55
u	32	9	65	14	80	0.91	0.68	1.01	0.32	1.83

Table 18. Chi-square tests for effect of similarity, with feature classes

a. Similarity, front/back			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	2.67	0.1023
Front	44 (*ii* + *ee*)	3 (*ie* + *ei*)	GoF	35.77	< 0.001 *[4]
Back	143 (*uu* + *oo*)	30 (*uo* + *ou*)	GoF	73.81	< 0.001 *
Totals	187	33	GoF	107.8	< 0.001 *

b. Similarity, high/mid			Test	$\chi^2(1)$	*P*
	Identical	Similar	CT	12.52	0.0004 *
High	116 (*ii* + *uu*)	51 (*iu* + *ui*)	GoF	25.3	< 0.001 *
Mid	71 (*ee* + *oo*)	7 (*eo* + *oe*)	GoF	52.51	< 0.001 *
Totals	187	58	GoF	67.92	< 0.001 *

Table 19. Chi-square tests for effect of similarity, with order classes

a. Similarity, front/back order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	5.11	0.0238 *
Front-Back	20 (*iu* + *eo*)	28 (*io* + *eu*)	GoF	1.33	0.2482
Back-Front	38 (*ui* + *oe*)	20 (*oi* + *ue*)	GoF	5.59	0.0181 *
Totals	58	48	GoF	0.94	0.3314

b. Similarity, high/mid order classes			Test	$\chi^2(1)$	*P*
	Similar	Dissimilar	CT	3.02	0.082
High-Mid	17 (*ie* + *uo*)	35 (*io* + *ue*)	GoF	6.23	0.012 *
Mid-High	16 (*ei* + *ou*)	13 (*oi* + *eu*)	GoF	0.31	0.5775
Totals	33	48	GoF	2.78	0.095

The results show significant overlap with the four daughter languages, with an interesting wrinkle. Identical V-Vs significantly outnumber similar V-Vs by about 6-to-1 when sorted by front/back classes (Table 18a) and 3-to-1 when sorted by height classes (Table 18b). There is an association with height classes in this last table, where the identical-to-similar ratio is much larger for mid vowels, but this is similar to what was found in Tongan (section 4.1). As for the differences in frequencies of similar and dissimilar V-Vs, the POC data shows the expected pattern when this comparison is sorted by high/mid orders (19b), but we actually find an important inequality in back-front orders (19a). Similar back-front V-Vs outnumber back-front dissimilar V-Vs by 2-to-1, a significant difference that may explain the same trends in Hawaiian and Fijian.

[4] This result should be taken with a grain of salt, because the value for front similar V-Vs drops below the threshold for a chi-square test, and *e* is rather rare in POC in general. Still, the fact that *e-e* has the highest overall frequency for stems with *e* in the initial syllable is telling.

The conclusion we can draw from these patterns is that the graded differences between identical, dissimilar, and similar V-Vs found in the daughter languages were present in Proto-Oceanic. One of the order effects was found, namely that dissimilar > similar in high-mid orders only, so it is reasonable to expect that this was continued in the daughter languages. Interestingly, the preference for back-front orders over front-back seems to be stronger in POC, so much so that we have a significant inequality going the opposite direction as expected. While this makes sense of the suggestive patterns in Hawaiian and Fijian, we did not find a dissimilar > similar inequality in front-back orders. Finally, these findings confirm our conjecture from section 3 that the sporadic total assimilations documented by Elbert in Proto-Polynesian are unlikely to be the primary cause of the vowel co-occurrence data documented here. They seem to have largely been inherited.

5.2 Proto-Malayo-Polynesian

We followed the same procedure as explained in section 2 and 5.1 for the PMP reconstructions in the online Austronesian Comparative Dictionary (see http://www.trussel2.com/acd/acd-pl_pmp.htm). There are 2,347 reconstructed words on this webpage, 1,902 of which represent unique stems, and of these, 1,703 are disyllabic. Table 20 shows the co-occurrence patterns of all vowel types and indicates the source of the mid vowels in the header with correspondences in Proto-Oceanic.

The analysis of vowel co-occurrence in Proto-Malayo-Polynesian (PMP) offers some value, but it is complicated by the lack of the mid vowels that we have investigated in the daughter languages and Proto-Oceanic. In particular, the only well-known source of *e* is PMP *ay. *o* has two sources, PMP **ə* and **aw*, but only **ə* has a relatively high occurrence in the available data. Furthermore, there are distributional restrictions that suggest some of the patterns can be excluded. First, **aw* and **ay* only occur in initial syllables a couple of times, and only as part of a reduplicated word, e.g., *hawhaw* 'to wash, rinse'. **uy* and **iw* never occur in initial syllables, and nine of the eleven occurrences occur after *a,* which does not concern us. We can greatly simply the chart below, therefore, by excluding **uy* and **iw* altogether, and initial **ay* and **aw*, as done in Table 21.

Table 20. Raw frequencies in Proto-Malayo-Polynesian disyllabic stems

POC	i	e	a	o	o	u	o	o
PMP	i	ay	a	ə	aw	u	uy	iw
i	103	2	114	41	8	84		
ay		1						
a	102	13	219	58	20	139	3	6
ə	55	5	98	139	5	80		
aw					2			
u	57	8	120	30	3	200	2	
uy								
iw								

Table 21. Raw frequencies in Proto-Malayo-Polynesian disyllabic stems, with exclusions

	i	ay (e)	a	ə (o)	aw (o)	u	i	ay (e)	a	ə (o)	aw (o)	u
i	103	2	114	41	8	84	1.57	0.35	1.00	0.74	1.08	0.81
a	102	13	219	58	20	139	0.99	1.43	1.23	0.67	1.72	0.85
ə (o)	55	5	98	139	5	80	0.77	0.80	0.79	2.31	0.62	0.71
u	57	8	120	30	3	200	0.73	1.16	0.89	0.46	0.34	1.62

Given these limitations on vowel co-occurrence, we cannot perform all of the chi-square tests we did above. For example, it is clear there is a difference beween *i-i* and *i-ay*, which became *i-e* (see their boxed O/E values in Table 21), but we do not have the same number of observations. We also note we do not know all of the sources of *e* in the daughter languages, so the low O/E of *i-ay* may be misleading. We can proceed with a bit more confidence with the the sources of back vowels, i.e., **ə, *aw,* and **u*, though we only have a handful of cases with **aw*. These seem to show the same patterns we have found in other languages: identical V-Vs are over-represented, and similar ones under-represented. Thus, if we sum the two available identical V-Vs (*ə-ə* and *u-u*) and compare that with the rest of the boxed back values in Table 21, we get 344 vs. 113 observations, which is a significant contrast (GoF chi-square = 116.764, $P < 0.001$). We also note that the same trend in order relative to height categories is observed, where high-mid *u-o* order is less frequent that the mid-high *o-u* order: 0.37 vs. 0.58, though this difference is less marked than we find in the daughter languages. To sum up, while the PMP vowel co-occurrence data is difficult to compare with the rest of the data, the data for sources of back vowels is certainly consistent with our findings in POC.

For the sake of comparison, these findings are also consistent with the findings in Chrétien (1965), who documents vowel co-occurrence of **i, *a, *ə,* and **u* in Proto-Austronesian (see his Table 11), using reconstructed vocabulary from Dempwolff (1938). Chrétien documented over-representation of identical vowels (using a different metric), which in his data have a mean O/E of 1.74; however, only *ə-ə*, with an O/E of 2.45, is exceedingly high, cf. 1.31 for *i-i* and 1.46 for *u-u.* Similar V-Vs are under-represented, with a mean O/E of 0.72, though there is not enough data to compare them with dissimilar V-Vs. The lowest O/E in Chrétien's dataset is 0.40 for *u-ə,* which is very close to the corresponding measurement above for PMP.

6. Summary and analytical focus

We recap the fundamental results in the summary table below, which shows mean O/E for the three V-V types and the results of all our chi-square tests, i.e., the goodness of fit (GoF) and contingency table (CT) tests; '*' indicates a significant test result.

Table 22. Summary of vowel co-occurrence data

	Samoan	Tongan	Hawaiian	Fijian	POC	PMP
$O/E_{Identical} > O/E_{Similar}$	2.06 > 0.94	1.91 > 0.91	1.86 > 1.10	2.27 > 0.86	3.20 > 0.68	1.83 > 0.94
GoF front	*	*	*	*	*	*?
GoF back	*	*	*	*	*	*?
CT front/back						*NA*
GoF high	*	*	*	*	*	*NA*
GoF mid	*	*	*	*	*	*NA*
CT height		*			*	*NA*
$O/E_{Dissimilar} > O/E_{Similar}$	0.94 > 0.58	0.91 > 0.68	1.10 > 0.68	0.86 > 0.53	0.68 > 0.48	0.94 > 0.60
GoF (all) front-back	*		*	*		*NA*
GoF back-front			(sim>dissim)	(sim>dissim)	*(sim>dissim)	*NA*
GoF high-mid	*	*	*	*	*	*NA*
GoF mid-high			*	*		*NA*

To generalize, all languages have significant over-representation of identical V-Vs relative to all other V-Vs, and this was a property of POC, so it was likely inherited by all the daughter languages. The observation that identical V-Vs outnumber similar V-Vs is usually found in all feature classes. However, in Tongan and POC, there is an association with vowel height (i.e., there is a significant CT test result).

All languages also have a contrast between dissimilar and similar V-Vs in high-mid orders, a distributional fact that also seems clearly inherited from POC. However, only some languages (Hawaiian and Fijian) have the contrast in the opposite mid-high order. Finally, most languages also have a contrast between dissimilar and similar V-Vs in front-back orders, and not the opposite order, but this does not seem to have been inherited. Interestingly, similar V-Vs in POC significantly outnumber dissimilar V-Vs in back-front orders, and we see the effects of this in Hawaiian and Fijian.

To sharpen the focus of the analysis, we are interested in the following core facts:

1. Basic preference in all languages for identical V-Vs over dissimilar and similar V-Vs, e.g., *i-i* > *i-e*.
2. Basic preference for dissimilar V-Vs over similar V-Vs in certain orders. In particular, all languages greatly prefer dissimilar over similar V-Vs in high-mid orders, e.g., *u-e* > *i-e*, and two of these languages (Hawaiian and Fijian) also prefer these V-Vs in the opposite order, as in *e-u* > *e-i*. Also, in three of daughter languages, dissimilar V-Vs are preferred to similar ones in front-back orders, as in *i-o* > *i-u*, but this preference is never found in the opposite order, **o-i* > *u-i*.

We can give a coarse-grained characterization of these generalizations by binning V-Vs into three well-formedness classes based on their O/E. Table 23 below shows how these generalizations play out in the four languages. The boxed regions under the language columns are created arbitrarily based on ±0.50 adjustments of O/E from 1.0, but they do correspond to many of the core generalizations we are interested in. Thus, all languages exhibit over-representation of identical V-Vs. Likewise, all languages have a 'normal' occurrence of dissimilar V-Vs and back-front and mid-high similar V-Vs (the features classes are shown the left, e.g., "FF" for front-front, "HM" for high-mid, etc.). There are two anomalies in Hawaiian, shown with arrows, which indicate over- or under-representation with respect to other V-Vs in the cluster. In the rest of the data, the Polynesian languages essentially differ from Fijian in where the cut-off is within the set of similar V-Vs. Tongan, Samoan, and Hawaiian are more

permissive, exerting a strong dispreference for high-mid similar V-Vs that agree in [back], but Fijian extends this ban to front-back V-Vs. The O/E values also reveal the order effects within the 'normally represented' V-Vs, as discussion in Samoan in section 2. Our analysis below examines individual languages and provides an analytical framework for ordering these V-Vs on a language particular basis.

Table 23. V-V orderings across languages; O/E values sorted by 1.50 (over-represented) > 1±0.50 > .50 (under-represented)

				Tongan	Samoan	Hawaiian	Fijian
Identical	FF	HH	i-i	1.87	2.18	1.75	1.95
	FF	MM	e-e	2.32	2.28	1.87	2.71
	BB	MM	o-o	1.85	1.92	1.70	2.02
	BB	HH	u-u	1.60	1.90	2.10	2.43
Dissimilar	FB	HM	i-o	1.25	1.19	1.73 (↑)	1.19
	BF	MH	o-i	0.84	0.81	0.95	0.78
	FB	MH	e-u	0.77	0.86	0.86	0.81
	BF	HM	u-e	0.77	0.93	0.87	0.66
Similar	BF	HH	u-i	1.17	0.97	1.28	0.76
	BF	MM	o-e	0.81	0.99	1.11	1.07
	FF	MH	e-i	0.64	0.68	0.29 (↓)	0.65
	BB	MH	o-u	0.72	0.54	0.69	0.53
	FB	HH	i-u	0.73	0.59	0.69	0.25
	FB	MM	e-o	0.66	0.50	0.85	0.31
	BB	HM	u-o	0.39	0.14	0.09	0.40
	FF	HM	i-e	0.32	0.28	0.29	0.27

7. Analysis: capturing relative well-formedness with parasitic harmony

In our analysis, we focus on the core facts of these systems: vowel co-occurrence is graded and it depends on similarity classes and order. The analysis below is intended address these problems, and also contribute to the large discussion of 'irregularity' in the analysis of vowel harmony. The research reviewed in the introduction argues that many kinds of irregularity must be integrated naturally into the analysis of the entire system. The fact that identical, dissimilar, and similar V-Vs classes seem to fall within a continuum calls for such an analysis as well.

The findings above show that the more similar two vowels are, the more likely they are to be subject to co-occurrence restrictions. Furthermore, within the set of similar V-Vs, certain order effects were found, e.g., high-mid orders were severely restricted. A number of contemporary theories of vowel harmony provide mechanisms for addressing these problems, and we draw on the theoretical insights of these works in our analysis. We follow the general approach in many theories of formalizing the general drive for harmony as a set of feature agreement constraints that ban disharmonic vowel combinations (Bakovic 2000; Finley 2009; Krämer 2003). A particularly important notion in our analysis is parasitic harmony, where harmony of feature F is predicated on agreement with feature G (Archangeli 1985; Cole 1987; Cole & Trigo 1989; Hare 1990; Jurgec 2011; Kaun 1995; Krämer 2003; Nevins 2010; Smolensky

2006; Wayment 2009). We do not have any particular commitments to the formalization of parasitic harmony, but note that Jurgec's Feature Spreading 2.0 model is well-suited to the empirical patterns we find here because of its parameterization of similarity and directionality effects.

The constraints given in Table 24 are from this theory, augmented with an AB/BA parameter for directionality. Feature Spreading 2.0 does not use this specific parameter, but we feel this theory is well-suited to the problem of directionality for the following reason. In Feature Spreading 2.0, each agreement constraint contains two parameters for adjacent root successive nodes X1 and X2, paraphrased as follows.

Assign a violation mark if:
(i) X1 is associated with Feature *F*, but X2 is not associated with *F*
(ii) if X1 is not associated with *F*, but X2 is associated with *F*.

Condition (i) requires agreement of X1 with X2, and (ii) requires agreement of X2 with X1. This is in essence a directionality effect: (i) is left-to-right agreement, and (ii) is right-to-left agreement. In our adapted system below, AB constraints require left-to-right agreement via (i), and BA constraints require right-to-left agreement (ii). It is important to note, however, that Feature Spreading 2.0 requires featural alignment, which we do not discuss here. In some ways, therefore, the present analysis is also similar to the directionality effects formalized in Hayes and Londe (2006), who treat back harmony in Hungarian as the result of *[+back][-back] to get the directionality effect.

Table 24. AGREE constraints, modelled after Feature Spreading 2.0 constraints

AGREEAB[αHIGH]-[+BACK]	If two vowels are associated with the same [αhigh], then if A is associated with [+back], then B must also be associated with [+back].
AGREEBA[αHIGH]-[+BACK]	If two vowels are associated with the same [αhigh], then if B is associated with [+back], then A must also be associated with [+back].
AGREEAB[αBACK]-[+HIGH]	If two vowels are associated with the same [αback], then if A is associated with [+high], then B must also be associated with [+high].
AGREEBA[αBACK]-[+HIGH]	If two vowels are associated with the same [αback], then if B is associated with [+high], then A must also be associated with [+high].
AGREEHIGH	Vowels agree in [αhigh].
AGREEBACK	Vowes agree in [αback].

With these four basic parasitic harmony constraints, and two general harmony constraints, it is a rather simple matter to establish the markedness relations implied by the frequency data with weight assignments from Harmonic Grammar (Pater 2009; Smolensky et al. 1992). With the weights given in Table 25 below, harmony scales rather well with the O/E values. There are many possible weight assignments that are consistent with the Samoan data, but these are chosen to give the largest amount of transparency within the tableau for the various effects of constraint violation.

Table 25. O/E scales by harmony, Samoan (compares with Tongan and Hawaiian)

		AGREEAB [HIGH]-BACK	AGREEBA [HIGH]-BACK	AGREEAB [BACK]-HIGH	AGREEBA [BACK]-HIGH	AGREEHIGH	AGREEBACK		
		-50	-250	-450	-250	-50	-50	*harmony*	O/E
identical	i-i							0	2.18
	e-e							0	2.28
	o-o							0	1.92
	u-u							0	1.9
dissimilar	i-o					1	1	-100	1.19
	o-i					1	1	-100	0.81
	e-u					1	1	-100	0.86
	u-e					1	1	-100	0.93
similar BF	u-i	1					1	-100	0.97
	o-e	1					1	-100	0.99
similar MH	e-i				1	1		-300	0.68
	o-u				1	1		-300	0.54
similar FB	i-u		1				1	-300	0.59
	e-o		1				1	-300	0.5
similar HM	u-o			1		1		-500	0.14
	i-e			1		1		-500	0.28

Because we are dealing with static phonotactic restrictions, we do not provide an analysis of the mappings from underlying forms to outputs. The differences in harmony values are posited as a way of structuring the data so that less frequent forms have lower overall harmony. In the literature on gradient phonotactics (see Hayes and Wilson (2008)), these differences in harmony are expected to correlate with differences in native speaker intuitions of gradient well-formedness. This predicts that native speakers should rate words with lower O/E values (and therefore lower harmony scores) as having lower levels of acceptability. While this is a question for future research, previous research has shown a connection between nonce-word ratings and O/E (Frisch et al. 2000). Nonetheless, we do conjecture a role for faithfulness constraints in input-to-output mappings to account for some of the oddball V-Vs patterns discussed below.

In addition to giving us the relative harmony values for classifying the data in Samoan, this analysis seems to extend well to the other languages. The same basic constraint system and weight assignments will apply to the data in Tongan, as the cut-offs in O/E values are very similar (see Table 23). Fijian, however, presents an interesting difference. Both high-mid and front-back V-Vs are severely under-represented in this language, unlike Samoan and Tongan. By increasing the negative weight of AGREEBA[αHIGH]-[+BACK] from -250 to -450, these two shunned V-Vs receive the same high harmony score of -500, as shown in Table 26 below.

Table 26. O/E scales by harmony, Fijian

		AGREEAB [HIGH]-BACK	AGREEBA [HIGH]-BACK	AGREEAB [BACK]-HIGH	AGREEBA [BACK]-HIGH	AGREEHIGH	AGREEBACK		
		-50	-450	-450	-250	-50	-50	*harmony*	O/E
identical	i-i							0	1.95
	e-e							0	2.71
	o-o							0	2.02
	u-u							0	2.43
dissimilar	i-o					1	1	-100	1.19
	o-i					1	1	-100	0.78
	e-u					1	1	-100	0.81
	u-e					1	1	-100	0.66
similar BF	u-i	1					1	-100	0.76
	o-e	1					1	-100	1.07
similar MH	e-i				1	1		-300	0.65
	o-u				1	1		-300	0.53
similar FB	i-u		1				1	-500	0.25
	e-o		1				1	-500	0.31
similar HM	u-o			1		1		-500	0.40
	i-e			1		1		-500	0.27

Hawaiian V-Vs are interesting in that they present some challenges to the basic system sketched in Table 25, but there seem to be tractable ways of extending it to Hawaiian as well. One immediate problem is that *e-i* is roughly equal to *i-e* in Hawaiian, which really sets Hawaiian apart from the other three daughter languages. With an O/E of .29, both V-Vs seem relegated to the lower ranks of the system. One idea for this case is to unify the constraints banning these V-Vs by removing the AB/BA parameter (or using the conjoined constraint from Jurgec (2011)), i.e., a general AGREE[αBACK]-[+HIGH] constraint, and assigning the unified constraint a higher negative weight. This move will have the effect of assigning greater markedness to *o-u*, but alternative tack may be applicable in this case. Another fact of interest in Hawaiian is the relatively high O/E of *i-o*, as well as the unusually low O/E of *u-o*. Indeed, *i-o* companies with identical V-Vs. While we have side-stepped the analysis of mappings so far, one idea would be that *i-o* has higher than expected O/E because it is the output of a mapping that takes *u-o* as input and results in *i-o*, i.e., fronting of *u* before a syllable containing *o*. It is not obvious what independent evidence there is for such a mapping, but it could give *i-o* the needed boost to account for these differences. In sum, Hawaiian seems to require some extensions of our core Samoan system, but ones that do seem tractable.

8. Conclusion

We have shown above how parasitic harmony constraints can provide the foundations of an analysis of similarity and directionality effects in the vowel co-occurrence patterns. How do the constraints in this analysis relate to tried-and-true constraints in other harmony systems? In other words, what is the appeal for these constraints in other Austronesian languages and more broadly? The constraints employed above are general constraints for parasitic vowel harmony, and we are essentially

arguing that the same constraints employed for [back], [high], and [round] parasitic harmony are active in Oceanic stem phonotactics. [back] harmony predicated on similarity in height, and height harmony predicated on sameness in [back] are routinely encountered in typologies of vowel harmony, and so they do not need to be invoked here. Indeed, we find evidence for the operative constraints within Austronesian: Northern Sarawak languages exhibit height harmony triggered by shared [back] specifications, effectively banning high-mid V-Vs and producing the effects similar to that predicted by AGREEAB[αBACK]-[+HIGH]. We are confident therefore that the constraints above relate the Oceanic data to other languages. However, given the problems documented above for Hawaiian, we do not yet know how far these cross-linguistic parallels can be drawn. For example, if the analysis of Hawaiian sketched above involving unfaithful mappings is unsuccessful, then we may have to result to constraints that target specific V-Vs, e.g., *e-i,* which starts to have less cross-linguistic appeal. We hope that future work documenting native speaker intuitions of vowel co-occurrence can help sort out whether these specific patterns can be resolved with more general or language particular constraints.

References

Alderete, John & Mark Bradshaw. 2013. Samoan root phonotactics: Digging deeper into the data. Linguistic Discovery 11.1-21.

Archangeli, Diana. 1985. Yokuts harmony: Evidence for coplanar representation in nonlinear phonology. Linguistic Inquiry 16.335-72.

Archangeli, Diana, Jeff Mielke & Douglas Pulleyblank. 2012a. From sequence frequencies to conditions on Bantu vowel harmony: Building a grammar from the ground up. McGill Working Papers in Linguistics 22.

Archangeli, Diana & Douglas Pulleyblank. 1994. Grounded Phonology Cambridge, Massachusetts: MIT Press.

Archangeli, Diana, Douglas Pulleyblank & Jeff Mielke. 2012b. Greater than noise: Frequency effects in Bantu height harmony. Phonological explorations: Emprical, theoretical and diachronic issues, ed. by B. Botma & R. Noske, 191-222. Berlin, Boston: De Gruyter.

Bakovic, Eric. 2000. Harmony, dominance, and control New Brunswick, NJ: Doctoral dissertation, Rutgers University.

Blust, Robert. 1970. *i* and *u* in the Austronesian languages. Working Papers in Linguistics 2.6, 113-45. Honolulu: Department of Linguistics, University of Hawaii.

—. 1996. Low vowel dissimilation in Ere. Oceanic Linguistics 35.96-112.

—. 2009/2013. The Austronesian languages Canberra: Pacific Linguistics.

Blust, Robert & Stephen Trussel. 2013. The Austronesian Comparative Dictionary: A work in progress. Oceanic Linguistics 52.493-523.

—. 2015 (ongoing). The Austronesian Comparative Dictionary, web edition, http://www.trussel2.com/acd/.

Busentiz, R. L & M. J Busenitz. 1991. Balantak phonology and morphophonemics. Studies in Sulawesi Linguistics, Part II. From NUSA, Linguistic Studies in Indonesian and Other Languages in Indonesia, ed. by J.N. Sneddon, 29-47. Jakarta: Atma Jaya Catholic University.

Capell, Arthur. 1941/1957/1968. A new Fijian dictionary Sydney: Australasian Medical Publishing Co. (1941 ed.).

Chomsky, Noam & Morris Halle. 1968. The sound pattern of English New York: Harper & Row.

Chrétien, C Douglas. 1965. The statistical structure of the Proto-Austronesian morph. Lingua 14.243-70.

Churchward, C. Maxwell. 1959. Tongan dictionary London: Oxford University Press.

Clements, G. N. & Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. The Structure of Phonological Representations, ed. by H.v.d. Hulst & N. Smith, 213-55. Dordrecht: Foris.

Cole, Jennifer. 1987. Planar Phonology and Morphology. Cambridge, MA: MIT Doctoral dissertation.

Cole, Jennifer & L. Trigo. 1989. Parasitic harmony. Features, Segmental Structure and Harmony Processes, ed. by H.v.d. Hulst & N. Smith, 19-38. Dordrecht: Foris.
Dempwolff, Otto. 1938. Vergleichende Lautlehre des austronesischen Wortschatzes. Supplement 3. Austronesisches Wörterverzeichnis Berlin: Reimer.
Elbert, Samuel H. 1953. Internal relationships of Polynesian languages and dialects. Southwestern Journal of Anthropology 9.147-73.
Finley, Sara. 2009. Formal and cognitive restrictions on vowel harmony: Johns Hopkins University Doctoral dissertation.
Frisch, Stefan. 1996. Similarity and frequency in phonology: Northwestern University Doctoral dissertation.
Frisch, Stefan A., Nathan R Large & David S. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. Journal of memory and language 42.481-96.
Goldsmith, John & Jason Riggle. 2012. Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. Natural Language and Linguistic Theory 30.859-96.
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.
Harrison, K. David. 1999. Vowel harmony and disharmony in Tuvan and Tofa. Paper presented at the Proceedings of 2nd Asian G.L.O.W.
Hayes, Bruce & Zsuzsa Cziráky Londe. 2006. Stochastic phonological knowledge: the case of Hungarian vowel harmony. Phonology 23.59-104.
Hayes, Bruce & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. Linguistic Inquiry 39.379-440.
Jurgec, Peter. 2011. Feature spreading 2.0: A unified theory of assimilation: University of Tromsø Doctoral dissertation.
Kaun, Abigail. 1995. The typology of rounding harmony: an Optimality Theoretic account: Ph.D. dissertation, UCLA.
Krämer, Martin. 2003. Vowel harmony and Correspondence Theory Berlin: Walter de Gruyter.
Kroeger, P. R. 1992. Vowel harmony systems in three Sabahan languages. Shifting patterns of language use in Borneo: Papers from the Second biennial Internation Conference, Kota Kinabalu, Sabah, Malaysia. Borneo Research Council Proceeding Series, Vol. 3, ed. by P.W. Martin, 279-96. Williamsburg, Virginia: Department of Anthropology, The College of William and Mary.
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.
—. 1971. The phonotactic structure of the morph in Polynesian languages. Language 47.
Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky. 1990. Can connectionism contribute to syntax? Harmonic Grammar, with an application. Proceedings of the 26th Regional Meeting of the Chicago Linguistic Society, ed. by M. Ziolkowski, M. Noske & K. Deaton, 237-52. Chicago: Chicago Linguistic Society.
Lynch, John. 2003. Low vowel dissimilation in Vanuatu language. Oceanic Linguistics 42.359-406.
Milner, George Bertram. 1966. Samoan dictionary: Samoan-English, English-Samoan London: Oxford University Press.
Nevins, Andrew. 2010. Locality in vowel harmony Cambridge, MA: The MIT Press.
Pater, Joe. 2009. Weighted constraints in generative linguistics. Cognitive Science 33.999-1035.
Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. NELS 23, 367-81.
Pukui, Mary Kawena & Samuel H. Elbert. 1986. Hawaiian Dictionary Honolulu: University of Hawaii Press.

Smolensky, Paul. 2006. Optimality in phonology II: Harmonic completeness, local constraint conjunction, and feature domain markedness. The harmonic mind: From neural computation to Optimality-Theoretic grammar. Vol. 2. Linguistic and philosophical implications, ed. by P. Smolensky & G. Legendre, 27-160. Cambridge, MA: The MIT Press.

Smolensky, Paul, Géraldine Legendre & Yoshiro Miyata. 1992. *Principles for an Integrated Connectionist/Symbolic Theory of Higher Cognition*. Computer Science Department, University of Colorado at Boulder Report CU-CS-600-92.

Vago, Robert. 1980. The sound pattern of Hungarian Washington, DC: Georgetown University Press.

Wayment, Adam. 2009. Assimilation as attraction: Computing distance, similarity, and locality in phonology: Johns Hopkins University.