GRADIENT MĀORI PHONOTACTICS[1]

Péter Rácz: *Department of Anthropology and Archaeology, University of Bristol <peter.racz@bristol.ac.uk>*

Jennifer Hay: *Department of Linguistics and New Zealand Institute of Language Brain and Behaviour, University of Canterbury <jen.hay@canterbury.ac.nz>*

Jeremy Needle: *Department of Linguistics, Northwestern University <jeremyneedle2011@u.northwestern.edu>*

Jeanette King: *Aotahi School of Māori and Indigenous Studies, and New Zealand Institute of Language Brain and Behaviour, University of Canterbury <j.king@canterbury.ac.nz>*

Janet B. Pierrehumbert: *University of Oxford e-Research Centre; New Zealand Institute of Language Brain and Behaviour, University of Canterbury; and Department of Linguistics, Northwestern University <janet.pierrehumbert@oerc.ox.ac.uk>*

Abstract

This paper provides a descriptive analysis of segmental distributions in the Māori lexicon. Focussing on the strict-CV subset of the lexicon, we examine co-occurrence restrictions of consonantal onsets and vowel nuclei of adjacent syllables. For consonants, we find that sequences that share the same place of articulation are under-represented. This shows a similarity avoidance effect in Māori, reported for other languages (Frisch et al., 2004; McCarthy, 1986). When we correct for the presence of reduplicants in the data-set, this under-representation includes sequences of identical consonants. Sequences of identical vowels are overrepresented, even when reduplicated syllables are taken into account. The results show that gradient phonotactic processes are operating in Māori beyond the categorical restrictions on syllable shape.

1. Introduction

Phonotactic constraints, as originally construed in classical phonological generative models, constitute categorical restrictions on patterns of phoneme occurrence and co-occurrence. English allows consonant codas, for example, but does not allow /h/ in coda position. Māori does not allow codas but - unlike English - allows /ŋ/ in initial position. It is now well documented, however, that languages can also contain non-categorical phonotactic restrictions which operate in a gradient fashion. For example, a commonly reported gradient pattern regarding consonants is the Obligatory Contour Principle for Place of Articulation (OCP-Place) – a dispreference for consonants sharing a place of articulation to occur in close proximity to each other. While the principle was first proposed to account for observed categorical constraints against such co-occurrences (McCarthy, 1986), a gradient tendency has also subsequently been observed in many languages (McCarthy, 1988; Berkley, 2000; Martin, 2007). Across a number of languages, words having highly similar consonants in close proximity are preferred less and come up in the lexicon less often. Frisch et al. (2004) argue that such patterns arise from a similarity-avoidance constraint in processing that disfavours repetition.

Not only are gradient phonotactic patterns evident in language, but they are accessible to native speakers. Many studies show that native speakers prefer nonce words that adhere well to the statistical patterns in their language (cf. e.g. Hay et al. 2004; Frisch et al. 2000; Bailey & Hahn 2001). These results provide further evidence that language-specific phonotactics are not solely a categorical part of linguistic competence but rather are gradient and reflect the statistics of the lexicon.

This paper is a short descriptive report describing several gradient phonotactic patterns in Māori.

In terms of categorical phonotactics, Māori is relatively simple. Syllables comprise a vowel nucleus, which can be short, long or diphthongal, and an optional simplex consonantal onset. These syllables freely combine with each other (cf. Harlow 2007; Bauer 2003). In this respect, it differs from most languages which allow vowel-only syllables, in that these are usually only permitted word-initially. Unlike most languages, Māori vowel-vowel sequences are not broken up by epenthesis (Blevins, 2008).

The probabilistic phonotactics of the language appear to be much more complex. Some indicative patterns have been provided in simple tabulations by Krupa (1968). He provides summary statistics for the segmental makeup of what he calls the possible set of Māori morphemes and words. Particularly relevant to the current paper, he notes that sequences of onsets where both onsets are either alveolar or labial (that is, have the same place of articulation) are less frequent than sequences with a different place. He also notes that some labial consonantal sequences are absent altogether. Krupa was writing before the term 'Obligatory Contour Principle' was first used in linguistics. Yet the tendency he notes is certainly consistent with the idea that a gradient OCP-type constraint is present in Māori. His analysis is restricted to two syllable words, and does not take into consideration the frequencies of the individual phonemes, and whether the apparent patterns of under-representation are statistically significantly different than would be expected if phonemes combined at chance.

Krupa's findings constitute an important starting point. Gradient OCP effects are reported from other languages, including Javanese, also an Austronesian language (Yip, 1988). Zuraw & Lu (2009) show that OCP effects in Austronesian languages are sensitive to morpheme boundaries. De Lacy (1997) finds OCP effects in Māori which mitigate against CV sequences in which both the C and V are labial. We are in fact studying the complementary case, relationships between CV sequences.

As noted in the literature on gradient OCP effects (McCarthy, 1986; Pierrehumbert, 1993; Frisch et al., 2004), pairs of identical consonants behave in two ways with respect to similarity avoidance. They might be avoided most markedly—as instances of maximum similarity. Alternatively, they might seem exempted from OCP altogether. Their prevalence in such cases can be seen as a result of copying processes in the morphology. This is, of course, a relevant point in the study of OCP in Māori, as the language has a number of reduplicative processes that can create pairs of identical sequences (Harlow, 2007).

Not all of these processes are necessarily transparent and active in the language. Blust (2007) convincingly argues that lexical bases in Austronesian languages are overwhelmingly bisyllabic. This means that longer Māori words—even if they are not the obvious result of an active morphological process—should be treated with the suspicion that they are historically morphologically complex.

This body of research points us to the following questions: Are there gradient restrictions on the co-occurrence of syllables in Māori? How are these restrictions affected by synchronic and diachronic morphological processes, which are likely responsible for most longer lexical entries as well as for sequences of identical consonantal codas and vowel nuclei?

In this paper, we present a simple statistical analysis of phonotactic patterns in the Māori lexicon. We examine co-occurrence restrictions between consonants and vowels in adjacent syllables. Our analysis confirms that adjacent pairs of syllables with homorganic non-identical onsets are avoided, occurring statistically less frequently than would be expected by chance. This occurs in all word positions, and is not restricted to two syllable words. The vowel analysis, on the other hand, shows that sequences of identical vowels are favoured. We also examine positional preferences for the distributions of different classes of segments, showing that both consonants and vowels show significant deviations from random in terms of the distribution of segments across different syllables in the word. We used R and Sweave (R Core Team, 2016; Leisch, 2002) for the processing, analysis, and write-up of the data.

2. Māori Phonotactics

2.1 Consonantal Patterns

The consonants of the language can be seen in Table 1 (cf. Krupa 1968; Harlow 2007). The labial fricative has varying realisations across speakers, with [f] being the dominant variant by a large extent (Maclagan & King, 2002). In this paper, we assume this realisation.

Table 1: Māori consonants.

	LABIAL	ALVEOLAR	VELAR	GLOTTAL
Plosive	p	t	k	
Fricative	f			h
Nasal	m	n	ŋ	
Tap		r		
Approximant	w			

We extracted a list of Māori lemmata from Williams (1957) archived online by Victoria University of Wellington. We extracted all headwords, and derived subwords that were subentries under the heads. For example, *ahu* is listed as a headword, and words such as *ahunga*, *whakaahu*, and *ahuahu* were listed as associated derived items.[2] All of these words were included in the initial lexicon. Māori contains a number of vowel sequences which may be pronounced as diphthongs, and also allows sequences of adjacent syllables which contain only vowels. Because a primary interest was in dependencies between adjacent syllables, we restricted our analysis to those cases where syllabification is unambiguous, and where we could be certain that the segments of interest were in adjacent syllables: namely, to words which do not contain any sequences of vowels, either as diphthongs or adjacent syllables. This subset of words provided a dataset of 8950 words. From this list we extracted a list of 22033 consonantal digrams (pairs of consonants) ignoring the intervening vowel (CVCV).

While morphological complexity undoubtedly affects the shape of these words, our main point of scrutiny is segmental patterns in the phonological word, and hence only took into consideration very broad and general patterns of word formation, such as reduplication (as will be described below). Figure 1 shows the distribution of consonants across different syllable positions in the word. As the syllable positions increase, we of course have diminishing observations. For example, all 6 syllable words also have a 2nd syllable, whereas the converse is not true. However, the rate of decrease and the overall distribution of observations differs across phonemes. Most saliently, the plosives (dashed lines) are over-represented towards the beginning of the word. The segment /r/ shows a unique profile. It is under-represented in initial position, but shows high rates of occurrence as onsets to the 2nd and 3rd syllables in the word. The segment /m/ appears to be over-represented at the beginning of the word, whereas the other nasals are underrepresented word-initially.

It should be noted that Krupa also briefly discussed positional effects - preferences for first or second syllable in his two syllable words. Based on simple tabulations, he observed that labials prefer to be in the initial syllable, alveolars prefer to be in the second syllable, and that velars do not display a preference. As we have seen, rather than place of articulation, the strongest overall effect appears to be on manner of articulation—with plosives set apart from other phonemes. In Krupa's tabulations, labials appear to prefer first position, but as we can now see, this apparent effect is not carried by all labials, and is likely driven by /m/ preferring to occur in initial position (Figure 1). Initial /m/ seems to be an exception to the overall tendency for labials to occur proportionately more later in the word than early in the word.

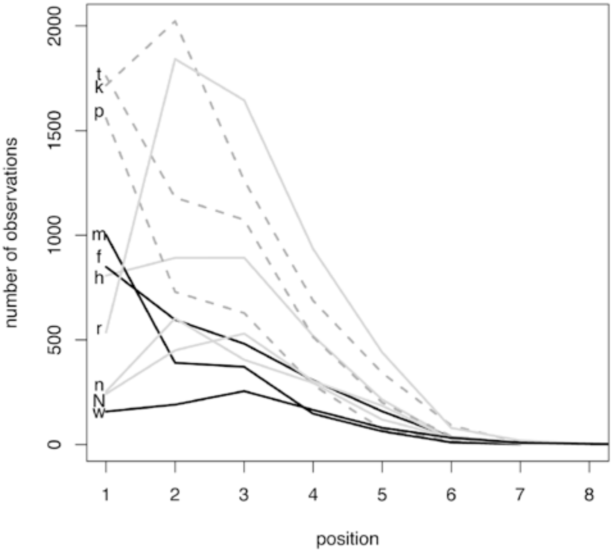

Figure 1: Positional distribution of consonants: stops (dashed) and labials (dark) each pattern together.

We next set out to test whether Māori has consistent restrictions on the co-occurrence of CV syllables as suggested by work on other languages (Frisch et al., 2004; McCarthy, 1986; Yip, 1988) and the patterns found by Krupa (1968).

The observed probability of each digram was calculated based on its frequency of occurrence in the corpus. Following Pierrehumbert (1993), this was contrasted with its expected probability, based on the observed frequency of the separate segments. More precisely, for a digram AB, the expected value was calculated as the observed frequency of A in position 1 of all digrams, multiplied by the observed frequency of B in position 2 of all digrams, divided by the total number of digrams observed. In this way, we could estimate the difference between observed and expected probability, by calculating the ratio of the observed value over the expected value. This Observed/Expected ratio (O/E) quantifies the degree to which a combination occurs more or less often than would be expected by chance. A value of 1 indicates that the combination occurs exactly with the frequency that we would expect, given the frequency of its parts. Numbers considerably smaller than 1 indicate under-representation, and numbers considerably larger than 1 indicate overrepresentation. The sequence f/k, for example, occurs much more often than one would expect by chance. In this case, this is no doubt due to the highly frequent prefix *whaka-*.

The Observed/Expected ratio is one measure of over- and under-representation. Below we use various, independent measures to recognise trends of over- and under-representation in the data.

Table 2: Observed over expected ratio for consonants in adjacent syllables.
The first segment in the combination is listed vertically, the second is listed horizontally. 0 reflects non-occurrence.

	p	m	f	w	t	n	r	k	N	h
p	**1.29**	0.15	0.71	0.34	1.04	1.21	1.12	1.13	0.65	0.98
m	0.13	**1.36**	0.47	0.21	1.78	1.51	0.92	0.87	0.96	1.04
f	0.05	0	**1.90**	0	1.20	0.73	0.68	2.71	0.48	0
w	0.06	0.10	0	**1.76**	1.10	1.57	1.52	0.91	0.16	1.19
t	1.25	1.09	1.38	1.01	**1.00**	0.54	0.98	0.95	1.25	0.94
n	0.90	1.15	1.77	2.35	0.54	**2.41**	0.13	1.24	0.58	1.62
r	1.40	1.38	1.28	1.32	0.77	0.30	**0.86**	0.83	1.76	1.21
k	1.32	1.42	0.95	1.19	0.97	1.03	1.18	**0.70**	0.39	1.12
N	0.31	0.73	1.22	0.76	1.59	0.82	1.24	0.28	**2.01**	1.18
h	0.82	1.10	0.14	1.41	0.60	1.55	1.12	1.05	1.49	**0.72**

Table 2 shows O/E values for all consonant pairs in adjacent syllables. Consideration of the diagonal reveals no particularly strong patterns of under- or over-representation indicating that there are no strong constraints regarding sequences of identical consonants, though sequences of identical sonorants are over-represented. The pattern in the table does suggest that sequences of non-identical consonants sharing a place of articulation are dispreferred. Indeed, if we look at the overall pattern of O/E as a function of homorganicity, we see that homorganic sequences are less likely overall. This becomes evident if we look at O/E values for consonantal place only (cf. Table 4).

Table 3: Observed over expected ratio for consonantal place in adjacent syllables, All consonants. The first segment in the combination is listed vertically, the second is listed horizontally.

	labial	coronal	velar
labial	0.66	1.20	0.98
coronal	1.36	0.84	1.10
velar	0.99	1.14	0.84

Table 4: Observed over expected ratio for consonantal place in adjacent syllables, without sequences of identical consonants. The first segment in the combination is listed vertically, the second is listed horizontally.

	labial	coronal	velar
labial	0.25	1.20	0.98
coronal	1.36	0.54	1.10
velar	0.99	1.14	0.33

O/E values for homorganic sequences (in which the two consonants share a place of articulation) are lower than for other sequences (Table 3). This becomes even more apparent if we exclude sequences of identical consonants (Table 4). We performed chi square tests on contingency tables for observed counts of homorganic sequences (separately for coronal, labial, and velar). We did this separately for all consonant counts and for counts excluding identical sequences. The patterns are highly significant in all cases. This, however, might be partly due to the large sample size. In order to see the effect size, we calculated the phi coefficient for each contingency table separately. The results are in Table 5.

Table 5: Strength of the underrepresentation effect for homorganic sequences, across consonantal place for all counts and for counts excluding identical sequences.

	ALL CONSONANTS	W/O IDENTICAL
coronal	-0.13	-0.30
labial	-0.14	-0.27
velar	-0.12	-0.29

Note that coronal, labial, and velar sequences are avoided to roughly the same degree. This overall pattern of underrepresentation does not change when we exclude sequences of identical consonants, although its effect does become much stronger.

In order to establish the statistical robustness of this observation, we stepwise fit a linear regression model on our consonant co-occurrence data using *observed over expected probability* (O/E) as the outcome variable, and the phonological features of the members of the consonantal digrams and whether these shared the same place as predictors. In this and subsequent models, we use the log of O/E, as it can have a tendency to have a long right tail. We exclude from the analysis any digrams which are not observed at all in the data-set.

Since, as we noted, digrams of identical consonants behave differently, we coded place as a factor with three levels, (i) no shared place, (ii) shared place, (iii) identical.

Since not all combinations of factors exist (a Māori consonant cannot be labial and velar at the same time, for instance), and since shared place and various phonological features are potentially collinear, we proceeded with bottom-up stepwise regression, testing the predictors individually and combining the significant ones if possible. The best model for the consonants is a simple one, retaining the feature of place, as shown in Table 6.

Table 6: Regression model of consonantal co-occurrence in CVCV sequences.

	Estimate	Std. Error	t value	Pr (>ltl)
Intercept (diff place)	0.04	0.07	0.57	0.57
identical	0.21	0.19	1.10	0.28
same place	-1.38	0.16	-8.92	<0.001

This confirms that sharing a place of articulation makes co-occurrence less likely. However, the repetition of the same consonant does not suffer a significant penalty. No significant interactions were found. This indicates that the dispreferrence for homorganic consonants is not strongly restricted to any particular class of sounds. We re-fit the model using, as the outcome variable, Observed minus Expected as an alternative measure of underrepresentation. The effect is similar.

In their study of Arabic, Frisch et al. (2004) demonstrate that two major factors determine consonantal co-occurrence: (i) whether the consonants are homorganic, i.e. share a place of articulation, as well as (ii) their general similarity, based on the number of natural classes they share. We used their measure of similarity (as implemented by Adam Albright, cf. Albright 2009) to calculate similarities between consonantal segments in Māori, and attempted to use this similarity measure in regression models, as above. However this approach did not deliver significant discrimination above the simple model reported in Table 3. This is chiefly, we believe, because the Māori consonantal inventory is relatively small and the number of homorganic segments is low.

Because our dataset includes words longer than two syllables, it is possible that the restriction on sequences of consonants sharing place of articulation is of different strength at different positions in the word. We thus calculated the observed/expected ratio for each consonant pair separately for each position in the word.

In Figure 2, we plot the observed counts of different digram types across specific positions. This plot faithfully shows the absolute values (which are diminishing). The plot shows lowess lines fit through relevant digrams. For example, at positions 1 and 2, the lines are fit through observation counts for 69 digrams not sharing place of articulation, 16 sharing place of articulation, and 10 which are identical. The number of digrams for which there are observations drops as we proceed through the word. Digrams for which we have no observations are not included in the plot. Thus, not only are the average number of observations per digram diminishing as we go through the word (as seen in the figure), so are the number of distinct digrams at each position (not visible from the figure). Consideration of the patterns indicates that the under-representation of adjacent consonants sharing a place of articulation is robust across all positions for which data is not scarce.

These results indicate the presence of co-occurrence restrictions that affect homorganic consonants but exempt identical ones. This is in line with the cross-linguistic pattern noted in Section 1 that co-occurrence restrictions can be upset by morphological processes that result in copying and reduplication.

Māori has a large number of word-formation processes that involve the reduplication of sequences of one or two syllables (Krupa, 1968; Harlow, 2007). Thus, while identical sequences appear not to be under-represented, it is likely that their frequency of occurrence is being bolstered by the inclusion of reduplicants in the data-set.

In order to see whether the observed patterns are an artefact of reduplication patterns, we created a restricted set of consonantal digrams. This set was based

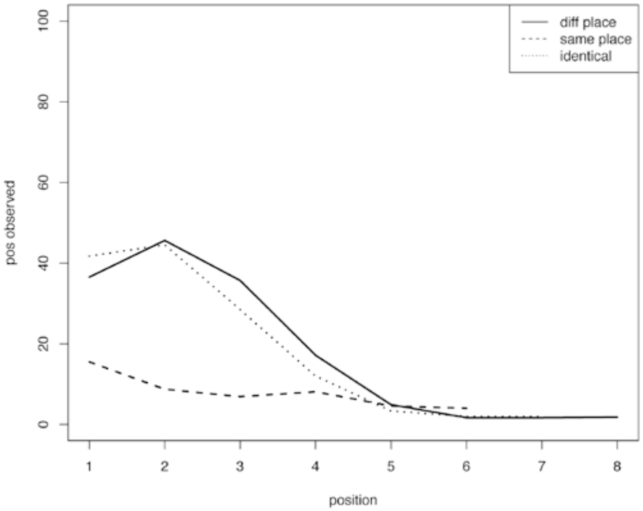

Figure 2: Observed/expected ratios of consonantal sequences across position in the word.

on only those words that had no direct repetition of syllables or sequences of syllables in them. This conservative approach to excluding effects of reduplication ensures that we can focus on patterns of co-occurrence that are independent from it.

This reduces the number of digrams under consideration from 22033, to 11854. The observed and expected values of digrams change considerably in the restricted set. Specifically, sequences of identical consonants become under-represented. However, homorganic sequences also remain under-represented relative to sequences of differing places of articulation, as can be seen in Figure 3.

This suggests that the pattern of Māori ‘suspending’ positional restrictions for sequences of identical consonants is only apparent - it is the artefact of the reduplication processes in the language. If we exclude words with reduplicated sequences, this pattern disappears and we arrive at a simpler tendency of broad homorganicity avoidance in onset sequences. This finding is consistent with earlier work on co-occurrence restrictions (cf. Section 1).

This becomes even clearer if we look at the observed over expected ratios of consonantal pairs in the restricted set (without reduplication) in Table 7.

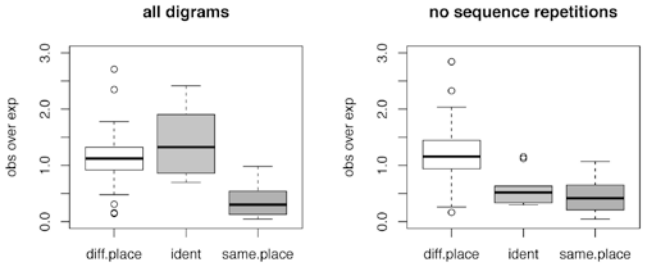

Figure 3: Positional distribution of consonants.

Regression analysis reveals that, in the restricted dataset, digrams of identical and homorganic consonants pattern together vis-à-vis other consonants (8). This is true both for distributions of observed over expected and observed minus expected counts. Reordering of factor levels reveals that the observed over expected ratio for identical and homorganic consonants significantly differs from the ratio for other consonants - their difference from each other, however, is only marginally significant (p=0.066).

Table 7: Observed over expected ratio for consonants in adjacent syllables.

Excluding words with reduplication patterns. The first segment in the combination is listed vertically, the second is listed horizontally. 0 reflects non-occurrence.

	p	m	f	w	t	n	r	k	N	h
p	**0.44**	0.19	0.93	0.43	1.01	1.36	1.20	1.26	0.65	1.02
m	0.21	**0.32**	0.66	0.30	2.01	1.69	0.84	0.86	0.94	1.15
f	0.05	0.00	**0.33**	0.00	1.30	0.75	0.73	2.84	0.38	0.00
w	0.00	0.21	0.00	**0.63**	1.06	0.97	1.71	0.93	0.17	1.50
t	1.33	1.20	1.43	0.94	**0.63**	0.61	1.07	0.99	1.13	1.05
n	0.84	1.35	1.82	2.33	0.64	**1.15**	0.10	1.42	0.91	1.70
r	1.69	1.53	1.29	1.45	0.87	0.42	**0.59**	0.75	2.04	1.21
k	1.50	1.61	1.11	1.28	1.00	0.93	1.25	**0.45**	0.41	1.19
N	0.50	0.55	1.54	1.37	1.85	1.07	1.05	0.35	**1.12**	1.14
h	0.95	1.16	0.26	1.42	0.64	1.77	1.15	1.04	1.63	**0.30**

Table 8: Regression model of consonantal co-occurrence in CVCV sequences.
Excluding words with reduplication patterns.

	Estimate	Std. Error	t value	Pr (>ltl)
Intercept (diff place)	0.11	0.07	1.63	0.11
identical	-0.73	0.18	-3.97	<0.001
same place	-1.14	0.15	-7.54	<0.001

Thus, Māori shows a clear effect of OCP, in which sequences of homorganic consonants are under-represented. This dispreference includes sequences of identical consonants. However, because of the high rate of reduplication in the lexicon, sequences of identical consonants are not under-represented in the lexicon as a whole. In other words - sequences of identical consonants are underrepresented, unless they are embedded in sequences of identical syllables.

2.2 Vowel patterns

We now turn to constraints on vowel position and vowel co-occurrence. Due to the orthographic conventions of Māori, adjacent sequences of vowels indicate either a diphthong or a vowel sequence. Since these are impossible to tell apart automatically, we again restricted our analysis to non-diphthongal nuclei of adjacent CV syllables (CVCV). This also includes long vowel nuclei. The Māori vowels are /i/, /e/, /a/, /o/, /u/, and their long equivalents.

Figure 4 shows the occurrence frequencies of the different segments across different syllable positions. As with the consonants, there is a reasonable amount of variation across distributions. /a/ is by far the most frequent vowel. Short vowels (dashed lines) are more frequent than long vowels (dotted lines). Long vowels are relatively rare beyond the second syllable. In initial position, back vowels and low vowels appear more frequent than front vowels and high vowels. /i/ and /e/ occur with higher frequency in the 2nd and third syllable than they do in the first.

In order to study co-occurrence patterns, we extracted digram frequencies of vowels in adjacent syllables in a way similar to the extraction of consonantal digrams discussed above. We calculated ratios of observed over expected values, and these are given in Table 9.

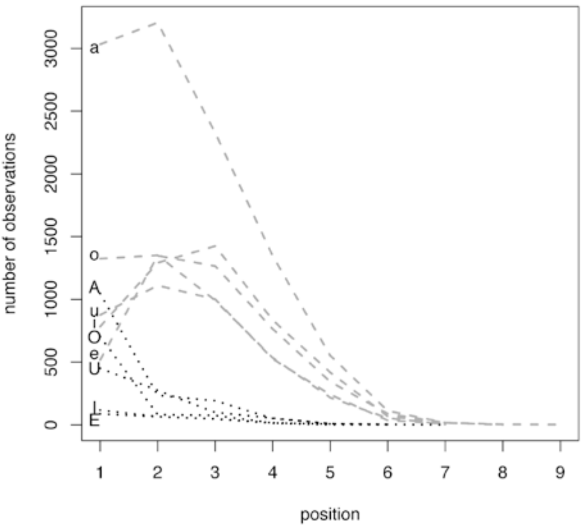

Figure 4: Positional distribution of vowels. Capital letters represent long vowels.

Table 9: Observed over expected ratio for vowels in adjacent syllables.
The first segment in the combination is listed vertically, the second is listed horizontally. 0 reflects non-occurrence. Capitalisation indicates a long vowel.

	i	I	e	E	o	O	u	U	a	A
i	**2.05**	0.86	0.45	0.12	1.33	0.53	0.47	0.40	0.85	0.64
I	0.75	**30.66**	0.65	2.19	1.11	2.46	0.00	0.00	1.06	2.07
e	0.49	0.31	**2.30**	0.60	1.00	0.36	0.78	0.62	0.79	0.28
E	0.32	0.00	1.78	**42.97**	0.12	3.12	0.32	0.00	0.80	1.50
o	1.03	1.15	1.18	0.65	**2.06**	0.38	0.77	0.69	0.61	0.41
O	0.92	4.92	0.95	3.87	0.79	**5.82**	1.10	2.83	0.73	2.51
u	0.84	0.20	0.82	0.26	0.39	0.15	**2.65**	0.60	0.89	0.52
U	0.38	1.03	1.02	2.64	0.89	2.48	0.81	**8.86**	0.94	4.03
a	0.93	0.51	0.75	0.76	0.69	0.98	0.88	0.71	**1.37**	0.80
A	0.85	1.24	0.82	1.19	0.83	2.99	0.90	2.50	0.92	**4.88**

Examination of the diagonal of this table reveals that identical vowels appear consistently over-represented in adjacent syllables. All sequences of adjacent identical vowels are more frequent than one would expect based on the frequencies of occurrence of the phonemes alone. Segments /i:/ and /e:/ are relatively rare in isolation. This fact contributes to the large O/E values for co-occurrence. They are quite rare phonemes, and thus the fact that they occur quite frequently in combination amounts to a large degree of statistical over-representation.

We stepwise fit a linear regression model on the vowel co-occurrence data *using observed over expected probability* as the outcome variable, and the phonological features of the members of the vowel digrams and whether these shared the same place and whether they were identical as predictors. The model returns a significant three-way interaction, indicating a significantly large degree of over-representation of sequences of adjacent identical long front vowels. The model shows that sequences of identical vowels are over-represented in general.

In order to determine whether the over-representation of identical vowels is related to reduplication, we excluded all words with repeated sequences from our material—restricting the set of word chunks the same way as we did for the consonants. This reduced the total vowel digrams under consideration from 22664 to 12409. We recalculated the O/E. The values are shown in Table 10.

This changes the profile quite considerably. While pairs of identical vowels still remain somewhat over-represented, it appears that the very strong tendency for identical vowels to occur in adjacent syllables was to a large degree carried by syllables that also share onsets. Notably, sequences of two /i:/s and two /e:/s are now unattested. Their over-representation in the larger data-set was driven entirely by words in which they occurred in adjacent identical syllables. This goes some way to explaining the previously observed interaction—in which sequences of identical long front vowels were over-represented.

Next we fit a simple linear regression model to the restricted data-set in Table 7, testing the phonological characteristics of the vowel sequences. The resulting model is shown in Table 11. It is very simple, revealing a remaining significant over-representation of sequences of identical vowels. We re-fit the model using observed minus expected as the outcome variable. That model shows the vowel identity factor to be highly significant.

Table 10: Observed over expected ratio for vowels in adjacent syllables, excluding repeated sequences.

The first segment in the combination is listed vertically, the second is listed horizontally. 0 reflects non-occurrence.

	i	I	e	E	o	O	u	U	a	A
i	**1.68**	1.55	0.56	0.23	1.45	0.88	0.46	0.61	0.90	0.78
I	1.02	**0.00**	0.97	0.00	1.21	3.77	0.00	0.00	1.12	2.51
e	0.57	0.53	**2.01**	1.02	1.03	0.71	0.94	1.07	0.86	0.27
E	0.33	0.00	2.33	**0.00**	0.22	3.38	0.46	0.00	1.18	1.69
o	1.04	1.34	1.25	1.04	**1.82**	0.60	0.86	0.73	0.65	0.52
O	0.94	4.74	0.83	4.61	1.08	**1.07**	1.15	3.02	0.70	2.76
u	0.92	0.35	1.01	0.23	0.39	0.24	**2.06**	0.96	0.95	0.71
U	0.39	1.75	1.11	3.40	1.01	2.76	0.94	**1.49**	0.90	4.60
a	1.00	0.61	0.78	0.98	0.74	1.09	0.93	0.84	**1.24**	0.87
A	0.82	1.38	0.78	0.67	0.80	2.49	0.99	2.23	1.07	**2.64**

Table 11: Regression model on the restricted set.

	Estimate	Std. Error	t value	Pr(>ltl)
Intercept (different segments)	-0.04	0.07	-0.51	0.61
identical segments	0.56	0.24	2.3	0.02

This means we have strong support for the overrepresentation of identical vowel sequences even if we remove all sequences of identical syllables.

3. Discussion

This paper builds on the simple tabulations provided by Krupa (1968), which examined some frequencies of occurrence and co-occurrence in two-syllable words. We extend this analysis to include longer words, and conduct statistical analysis to identify the degree to which co-occurrence patterns deviate significantly from chance.

Our analysis reveals consistent and statistically significant deviations from chance in the co-occurrence of Māori consonants and vowels. Adjacent onsets sharing the same place of articulation are avoided. Adjacent identical vowels are overrepresented. Both vowels and consonants show positional effects in terms of their overall distributions across syllables.

The consonant analysis confirms that the trends reported in Krupa (1968) are robust, and do appear to be a manifestation of an OCP-type pattern. This closely compares to the gradient avoidance of similarity observed in other languages (Frisch et al., 2004; Martin, 2007). By way of explanation for such patterns, Frisch et al. (2004) point to work on the effect of repetition on speech production and speech processing, showing that the repetition of segments is taxing to language processing and increases the chance of parsing errors. Frisch and his colleagues show that gradient similarity avoidance in the lexicon (wherein sequences of similar segments occur less often than they would if co-occurrence were based on chance) is relatively common, citing the example of Pierrehumbert (2006), who demonstrates that triconsonantal clusters in English morphemes are more restricted than similar clusters on morpheme boundaries. McCarthy (1986) and Frisch et al. (2004) discuss the fact that languages vary in with respect to the treatment of identical consonants. Totally identical consonants are sometimes permitted by languages displaying OCP-constraints, and sometimes excluded. In Māori, adjacent identical consonants are under-represented in the monomorphemic lexicon. However, they are well represented when sequences of identical syllables are included.

If CV syllables do not combine freely, as our results indicate, this could also be helpful for word segmentation, since the transitional probabilities of syllables within and across word boundaries will be different (Harris, 1955), a cue which adult listeners can exploit to locate word boundaries (cf. e.g. Saffran et al. 1996; Cairns et al. 1997). Many segmentation algorithms rely on identification of low probability diphones. Our results however suggest that, if Māori listeners use phonotactic patterns to segment the speech stream at all, then the relevant patterns extend over greater distances than the diphone.

This simple examination reveals that Māori phonotactics are more complex than what the simple syllable structure of the language might imply, and that a number of significant statistical tendencies underlie the phonological grammar. We have not yet examined longer-distance dependencies, nor whether there are patterned constraints that occur within the syllable. Considerable future work also awaits with respect to word shapes that contain sequences of multiple vowels.

According to our analysis, Māori restricts the co-occurrence of homorganic codas, with the exception of identical ones. As such, it fits into the cross-linguistic typology of co-occurrence restrictions (McCarthy, 1988). The exception of identical consonants disappears in a subset of the vocabulary that excludes words containing reduplicated sequences. This supports the assumption that the apparent exceptionality of identical consonants results from morphological processes in the language (cf. e.g. Yip 1995).

What remains clear is that while Māori phonotactics may appear simple on the surface, a number of gradient patterns and restrictions work together to shape the Māori lexicon. An interesting question for future experimental work is how much implicit knowledge of these gradient patterns speakers of Māori might have.

Note

[1] This project was made possible through the support of a subaward under a grant to Northwestern University from the John Templeton Foundation (Award ID 36617), and a grant from the Royal Society of New Zealand Marsden Fund (UOC1502). The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of the John Templeton Foundation or the Royal Society of New Zealand.

[2] ‘Ahu’, *a heap*; ‘Ahunga’, *heaping up*; ‘Whakaahu’, *to make a heap*; ‘ahuahu’, *to heap*.

References

Albright, Adam. 2009. Feature-based generalisation as a source of gradient acceptability. *Phonology* 26(01). 9–41.

Bailey, Todd M. & Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44(4). 568–591.

Bauer, Winifred. 2003. Māori. London: Routledge.

Berkley, Deborah Milam. 2000. *Gradient obligatory contour principle effects*. PhD Dissertation, Northwestern University .

Blevins, Juliette. 2008. Consonant epenthesis: natural and unnatural histories. In Good, Jeff. (Ed). *Linguistic Universals and Language Change*, Oxford: Oxford University Press. 79–107.

Blust, Robert. 2007. Disyllabic attractors and anti-antigemination in Austronesian sound change. *Phonology* 24(01). 1–36.

Cairns, Paul, Richard Shillcock, Nick Chater & Joe Levy. 1997. Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology* 33. 111–153.

De Lacy, Paul. 1997. A co-occurrence restriction in Māori. *Te Reo* 10–44.

Frisch, Stefan A., Nathan R. Large & David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42(4). 481–496.

Frisch, Stefan A., Janet B. Pierrehumbert & Michael B. Broe. 2004. Similarity avoidance and the OCP. *Natural Language & Linguistic Theory* 22(1). 179–228.

Harlow, Ray. 2007. *Māori: A linguistic introduction*. Cambridge: Cambridge University Press.

Harris, Zeilig S. 1955. Phoneme to morpheme. *Language* 31. 190–222.

Hay, Jennifer, Janet Pierrehumbert & Mary Beckman. 2004. Speech perception, well-formedness, and the statistics of the lexicon. In Local, John, Ricard Ogden, & Ros Temple (Eds.), *Phonetic Interpretation: Papers in Laboratory Phonology VI*. Cambridge: Cambridge University Press. 58–74.

Krupa, Viktor. 1968. *The Māori language*. Nauka.

Leisch, Friedrich. 2002. Sweave: Dynamic generation of statistical reports using literate data analysis. In: Härdle W., Rönz B. (eds) *Compstat*. Physica, Heidelberg. 575–580.

Maclagan, Margaret & Jeanette King. 2002. The pronunciation of wh in Māori: A case study from the late nineteenth century. *Te Reo* 45. 45–64.

Martin, Andrew Thomas. 2007. *The evolving lexicon.* PhD dissertation, University of California Los Angeles.

McCarthy, John J. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 51. 207–263.

McCarthy, John J. 1988. Feature geometry and dependency: A review. *Phonetica* 45(2-4). 84–108.

Pierrehumbert, Janet B. 1993. Dissimilarity in the Arabic verbal roots. *In Proceedings of NELS*, vol. 23, 367–381.

Pierrehumbert, Janet B. 2006. Syllable structure and word structure: a study of triconsonantal clusters in English. In P. Keating (Ed.), *Phonological Structure and Phonetic Form*. Cambridge: Cambridge University Press. 168-188.

R Core Team. 2016. R: A language and environment for statistical computing. R Foundation for Statistical Computing Vienna, Austria. https://www.R-project.org/.

Saffran, Jenny R., Elissa L. Newport & Richard N. Aslin. 1996. Word segmentation: the role of distributional cues. *Journal of Memory and Language* 35. 606–621.

Williams, Herbert W. 1957. *A dictionary of the Māori language*. R E Owen, Government Printer, Wellington, New Zealand.

Yip, Moira. 1988. The obligatory contour principle and phonological rules: A loss of identity. *Linguistic Inquiry* 19(1). 65–100.

Yip, Moira. 1995. Repetition and its avoidance: The case in Javanese. Department of Linguistics, University of Arizona (Tucson, AZ) .

Zuraw, Kie & Yu-An Lu. 2009. Diverse repairs for multiple labial consonants. *Natural Language & Linguistic Theory* 27(1). 197–224.