An analysis of type inflation and error in the Legal Māori corpus

Contents

The Legal Māori Corpus spans some of the earliest writing in Māori to contemporary legal and political texts. These texts have been divided into three corpora based on the period in which they were written: pre-1910, 1910-1969 and post-1970. The pre-1910 corpus was compiled using printed texts digitised by the New Zealand Electronic Text Centre (Darwin & Stephens, 2009), while the post-1970 corpus consists of a mix of texts that were 'born digital' and were retrieved from online publications, and other published texts which were rekeyed by typists with expertise in digitising texts but without knowledge of te reo Māori. The corpora contain the largest structured digital collection of written Māori, with around 8 million words, but they also contain a large number of errors and variant forms of words. These errors and variant words mean that the total wordlist for the corpora yields around 60,000 unique words (known as types), which is considerably more types than one might expect to find. This report summarises the causes of error and other forms of type inflation (where one term is represented by a number of different forms, not as a result of error but as a result of orthographic, dialectal or lexical variation). Errors in the corpus include both original spelling errors, particularly in the variable use of macrons, and errors which were introduced in the digitisation and rekeying processes. Type inflation from language variation is due to the use of different orthographic conventions (chiefly differences in the use of macrons, double vowels and unmarked long vowels), and to variations in the language used by speakers of modern Māori. This variation can be attributed to both the use of reo-a-iwi, or dialectal differences in te reo Māori and also to variations in usage that reflect the development of modern Māori.

This report focusses primarily on the post-1970 corpus because a more detailed study was made of these texts in order to evaluate the accuracy of the rekeying process. This corpus of post-1970 texts in the legal Māori corpus contains just under two million words (tokens), which make up 23,457 unique types across 577 texts. However, this report also gives an overview of error and type inflation in the two earlier corpora in order to give a picture of the rate of error and other forms of type inflation in the whole corpus. The error analysis was done using the Māori-only portion of the Legal Māori Corpus[1] except where indicated.

Error analysis

Error identification was undertaken for two main reasons. Firstly, error identification is important for ascertaining the nature of the errors in the corpus by assessing the frequency of errors, the processes that caused errors to be introduced into the corpus and whether particular texts were more prone to error than others. Secondly, error identification is necessary in order to improve the quality of the wordlist by estimating the number of meaningful types in the corpus and identifying likely 'junk' types.

In order to assess the nature of error in both the corpus and the wordlist, two processes of error identification were used. The first approach was to take a sample of high-frequency words which were likely to be prone to error and to identify the errors associated with these words in order to identify patterns in errors and texts which were particularly prone to error. This process was directed at identifying patterns of error in the corpus. The second process was to scan the wordlist and identify obvious errors (junk types) for removal in order to clean up the wordlist.

Sampling

Initial error analysis was concentrated on the pre-1910 corpus because the high number of types in this corpus, and in particular the high number of types that occurred only once in the corpus, indicated that it was likely this corpus contained a high number of errors. Analysis of the pre-1910 corpus indicated that these errors typically took two forms. The first of these was letter substitution errors, where one letter had been swapped with another, for example *iēnei* rather than *tēnei*. The second area of focus was word boundary errors, for example *itēneiwā* rather than *i tēnei wā*. In order to analyse the patterns of error in the post-1970 corpus, the researcher scanned the wordlist to ascertain whether errors in the post-1970 texts also tended to be letter substitution and word boundary errors. Reading the wordlist in the second stage of the error identification process confirmed that these two patterns accounted for the majority of error in the corpus, and thus it was possible to look for errors in the post-1970 corpus using the same sampling strategies that had been used to identify errors in the pre-1910 corpus.

In order to identify likely errors in the post-1970 corpus, I began by identifying high frequency words of five or more letters on the grounds that there were likely to be more examples of misspellings of high frequency terms. Restricting the search terms to words of five or more letters excluded the

[1] The full Legal Māori Corpus contains texts that include both English and Māori. The corpus was subsequently marked up to distinguish the two languages. This analysis was done using the Māori-only portion of the corpus except for when investigating error types that were removed as part of the process of marking up the English portion of the corpus. This report also comments on the small number of English types in the Māori-only corpus which were not correctly identified in the marking up process.

large number of two-syllable words which differ from other legitimate types in the corpus by one letter, or which form shorter parts of longer legitimate words. For example, the highest-frequency four-letter term in the post-1970 corpus is *mahi*; searching for letter variations of *mahi* returns 27 types, of which four appear to be erroneous spellings of *mahi* and 23 are either legitimate words or erroneous spellings of other words. In contrast, searching for letter variations of the highest-frequency five-letter term *Māori* returns seven types, of which three are erroneous spellings, one is the alternate form *Maori* with no macron and three are either legitimate words or erroneous spellings of other words. Working with terms of five letters or more thus results in a much more concentrated collection of errors.

I created a word list of the post-1970 corpus using WordSmith Tools version 5.0 (Scott, 2010), then identified the highest frequency words of five letters or longer using the *length* function in Excel. The 100 highest-frequency words of five or more letters were then inputted into the customised Excel macro Wordsearch[2]. For example, for the search term *kaupapa,* Wordsearch returns the following letter variations: *raupapa, haupapa, kuupapa, kaupopa, kaurapa, koupapa* and *naupapa,* along with their relative positions in the word list of the corpus. The word list is ordered by frequency so words which appear more often in the corpus have a higher ranking. In this case, the search term *kaupapa* returns:

KAUPAPA (Letter Variation)	Rank number	Frequency
RAUPAPA	1408	73
HAUPAPA	5296	7
KUUPAPA	5837	6
KAUPOPA	15192	1
KAURAPA	15196	1
KOUPAPA	15796	1
NAUPAPA	16868	1

We can assume that high-frequency words are almost certainly genuine words rather than errors. For example, the term *raupapa* is ranked at 1408 in the wordlist and appears 73 times in the corpus. Similarly, *haupapa* and *kuupapa* are also genuine words. The remaining four terms each appear once only in the corpus. The next step is to ascertain whether these words are legitimate, low-frequency words or errors. To do this, the context in which these words appear is examined using the *concord* function in Wordsmith and the terms may also be cross-checked against existing dictionaries of Māori. Williams' (2004) dictionary gives an entry for *kaurapa* that is consistent with the context in which this word appears but the remaining three terms *kaupopa, koupapa* and *naupapa* all appear in contexts where *kaupapa* would make more sense, and Williams does not give entries for these words. From this, it may be assumed that these words are errors. Where it was not certain whether the terms returned were errors, a spot check was made of the original sentence the term occurred in using the Concord function of WordSmith Tools.

[2] Wordsearch is an Excel macro developed by Neal Osborne for the Legal Māori Project in order to assist in identifying letter substitution and word boundary errors. The macro compares words in a search list against the whole wordlist and identifies words which differ from the search terms by one character (letter variation function) and words which contain the search word (substring function).

Similarly, if the search term *kaupapa* is searched using the substring function, the following results are returned.

KAUPAPA (Substring)	Rank number	Frequency
KAUPAPAHERE	1097	106
WHAKAKAUPAPATANGA	1865	46
WHAKAKAUPAPA	2318	32
WHAKAKAUPAPATIA	3150	19
KAUPAPATIA	7171	4
KAUPAPAO	10265	2
NGĀKAUPAPA	10830	2
TEKAUPAPA	11932	2
TĒNEIKAUPAPA	11945	2
WHAKAKAUPAPAHIA	12532	2
ĀKAUPAPA	13022	1
KAUPAPAAROTURUKI	15180	1
KAUPAPAHEREATE	15181	1
KAUPAPAPA	15182	1
KAUPAPAPAHERE	15183	1
KAUPAPATANGA	15184	1
KURAKAUPAPA	15869	1
RAKAUPAPA	18702	1
WĀHANGAKAUPAPA	21505	1
WHAKAKAUPAPANGIA	22141	1
WHAKAKAUPAPATIAHEI	22142	1

Again, the higher frequency terms *kaupapahere, whakakaupapatanga, whakakaupapa, whakakaupapatia* and *kaupapa* can immediately be identified as legitimate types and the single occurrence *kaupapatanga* also falls into this category. Of the remaining terms *kaupapao, ngākaupapa, tekaupapa* and *tēneikaupapa* appear twice each in the corpus and can clearly be identified as word boundary errors between *kaupapa* and the grammatical words *o, ngā, te* and *tēnei* respectively. The single occurrences *ākaupapa* and *kaupapahereate, rakaupapa* and *whakakaupapatiahei* are likewise word boundary errors of *ā kaupapa*, *kaupapahere a te*, *ra kaupapa* and *whakakaupapatia hei.* Although they are not word boundary errors, *kaupapapa* and *kaupapapahere* also are clearly erroneous words containing a repeated –*pa*- syllable.

The remaining words demonstrate some of the difficulties in deciding what to classify as errors. *Whakakaupapahia*, appearing twice in the corpus, and *whakakaupapangia*, appearing once, contain alternative passive endings for *whakakaupapatia*. While these variations are clearly used less frequently than *whakakaupapatia* (19 occurrences across 16 texts), they are most likely are a faithful representation of the language used by speakers of te reo Māori. This variation bears out Harlow's (2007, pp. 115-118) observation that although –*tia* is the most productive of the passive suffixes, being most frequently used as the default suffix with transitivised verbs such as *whakakaupapa* or to transliterated words, some Eastern dialects use –*ngia* or –*hia* for this purpose. There are also a small number of words where different passive or nominal suffixes are used for different senses of a word, such as the nominalisations of *ako*: *ākonga* (student), *akomanga* (class, classroom) or *akoranga* (subject, course of learning). For these reason, I have taken a conservative approach to identifying error and have excluded low-frequency passive and nominal endings from the error list. Similarly, where there is confusion about whether a type may represent a word boundary error or a compound word, I have not categorised the type as an error. The search for *kaupapa* returned one occurrence each of the terms *kaupapaaroturuki, kurakaupapa, wāhangakaupap*a. While these are clearly not compound words in common usage, it is not possible to say that the original author did not intend to treat these as compound words. Harlow's (2007, pp. 129-131) indicate several orthographic patterns used in forming compound words including one word with no spacing, hyphenated words and two words. In light of this orthographic diversity, I have not treated as errors those terms which could conceivably be compound words of the noun+modifier variety. It is difficult, even with reference to the original text, to determine the authors' intent in many cases and this falls outside the scope of estimating the type count in the corpus. The user of the wordlist should therefore be aware that the wordlist contains a large number of types which are not in common usage but which can be considered true types rather than errors.

Searching for letter variation associated with sample search words (the 100 highest frequency terms of five letters or more) and sorting for errors returned 345 letter substitution errors, of which there were 205 unique errors. This process also returned 537 word boundary errors, of which 412 were unique errors. The letter variation errors were sorted by the original, intended letter and the error letter. This process showed that most of the errors appeared to result from confusion between similar-looking letters. The most common errors in the sample were:

57	a to o
41	t to r
34	a to e
29	o to a
26	o to e
23	e to a

Scanning the wordlist for error

The second step in the process of error identification process was to scan the wordlist for types that were clearly erroneous. Limitations of this process were: looking at the words out of context, the limited vocabulary of the reviewer and the large number of types to be reviewed (~23,500).

The reviewer identified a list of 2,409 errors including misspelled words, run-on words, macron errors and English words. It is likely that if more detailed analysis of the wordlist were undertaken, more errors would be found. As it stands, these types represent around 10% of the number of types in the overall corpus. The overall scan of the wordlist confirmed that looking for one-letter variations and word boundary errors was a sound process because these types accounted for the majority of errors identified. Other errors included a small number of errors involving missing letters or letters that had been added in (as in *whakpapa* for *whakapapa* and *enngari* for *engari* respectively) or occasionally letters that had been swapped around.

English

There were 225 English types identified in the Māori-only corpus. Most of these are concentrated in a few texts and so it should be a straightforward matter to correct the mark-up process to exclude these passages from the Māori-only corpus. However, a smaller number of types point to problems distinguishing English and Māori in the mark up process. For example, when I investigated the type *who*, it became apparent that the process of marking up English for removal did not remove English words that could superficially appear to be Māori. These were words which contained only letters that appeared in the Māori alphabet and which contained only open (consonant-vowel or vowel only) syllables.

For example, the file SetTeUrioHauHCS2000M.txt contained the following text in the English-retained corpus:

> The Crown acknowledges that the operation and impact of the Native land laws (including the laws governing the operation of the Validation Court) had **a** prejudicial effect on those of **Te Uri o Hau who** wished **to** retain their land and that this was **a** breach of **Te Tiriti o Waitangi/**the Treaty of **Waitangi** and its principles. The Crown also acknowledges that the awarding of reserves exclusively **to** (Bold type added for emphasis)

And the following text in the Māori-only corpus:

> a Te Uri o Hau who to a Te Tiriti o Waitangi/ Waitangi to Te Uri o Hau to a Te Uri o Hau;

The remaining text in the Māori-only corpus indicates that the process of removing English from the texts sometimes results in residual gibberish. Users of the text who encounter such passages may wish to compare the texts they are working with against the corpus with English retained. A further implication of this is that the type/token ratio for certain types which occur in both English and Māori (such as *a* and *to*) may be distorted in the corpus, although because these are very high-frequency words it is not practical to determine to what extent this has occurred; *a,* for example, appears in the top ten high-frequency words in both te reo Māori and English (Boyce, 2006; Kilgarriff, 2006).

Origins of the errors

An initial look at the distribution of errors in the corpus indicates that some texts have a much higher concentration of errors than others. Furthermore, the kinds of errors identified in the sampling process showed that many of the errors were likely to have been introduced into the corpus in the digitisation process. Reasons for believing this are that the common letter substitution errors appear to concentrated around letters that are visually similar such as *a* and *o*, and *t* and *r*. Furthermore, high frequency words are not words which one would expect a speaker of te reo to misspell, yet there are quite a large number of errors associated with these high frequency terms.

A sample of texts that contained a high numbers of errors was selected for checking. The sample contained texts that were identified as having a very high number of errors and texts that were representative of the different categories of texts in the corpus, including texts that had been scanned and digitised, texts that were originally published in .pdf format and Hansard records retrieved from web pages. Analysis of this sample confirmed that many of the errors identified were present in the digitised texts but *not* in the original scanned texts. Many word boundary errors corresponded with line breaks in the original texts, where the original text gave two words and the digitised text combined the last word of one line with the first word of the next. Letter variation errors also appear to have been introduced in the digitisation but there were no clearly identifiable factors that led to these errors. It does seem that a small number of these errors were introduced when the scanned document was of poorer quality, but this does not seem to account for the majority of the letter variation errors. It is possible that the error was introduced when the texts were rekeyed by typists who were unfamiliar with te reo Māori. An exception to this pattern of error introduced in the digitisation process was the error profile in the Hansard texts. The word boundary errors in the Hansard texts were also found in the online publications of these speeches and the reasons for the errors in these texts are unclear.

The corpus with English retained

The majority of the error analysis was conducted using the Māori-only corpus with English removed. However, examining the full wordlist containing types that were identified as both English and Māori revealed that there were a large number of error types which had been removed in the process of marking up the languages in the corpus. In order to determine the extent of this issue, the wordlist of the Māori-only corpus was stoplisted against the full corpus so that only the terms which were not found in the Māori-only corpus were retained. This list contained 18,636 types, the majority of which were English words. However, 1,813 words or 10% of the types were identified as potentially being Māori words, either errors or words which had been eliminated from the Māori-only corpus because, for example, they were personal names which appeared in a passage of text in English. Errors were predominantly letter variation (e.g. *whakatlka* for *whakatikia*) or missing letter errors (e.g. *whakhaere* for *whakahaere*) that meant the resulting error type violated one of the phonetic rules of Māori and were consequently eliminated as being English types. Also eliminated were words with medial upper case letters, which occurred in word boundary errors involving proper nouns, such as *anaTe Karauna* for *ana Te Karauna* or *NgaiMāorie* for *Ngai Māori e*. It is unclear as to why these words appear to have been flagged as English in the mark-up process. In all, words which violated the phonetic rules of Māori (letter variation and missing letter errors) accounted for 794 of the possible error types. These additional error types would take the total number of errors identified in the process of scanning the post-1970 wordlist to 3,203, and represent 25% of the error types identified in the corpus.

Orthographic variation

There are three systems used to represent long vowels in Māori: macrons, double vowels and unmarked long vowels. Vowel marking systems have changed over time; long vowels are typically unmarked in earlier texts but macrons predominate in contemporary texts and in their orthographic guidelines, Te Taura Whiri treats macrons as the “established means of indicating a long vowel” (2012, p. 4). However, this convention is not universal and many writers of texts in the post-1970 corpus use unmarked long vowels, which contributes significantly to the high number of types in the corpus. This is because the corpus contains both the macronised and unmarked variants of the same word, for example, *Māori* and *Maori*. Double vowels (as in *Maaori*) are used by only a few writers in the contemporary corpus but still contribute somewhat to the large number of types. In order to understand to what degree alternative long vowel orthographies lead to inflation of the type count in the corpus, we can start firstly by looking at the distribution of the different vowel marking systems in the corpus. Then we can estimate the number of additional types in the corpus that occur as a result of alternative orthographic systems.

The great majority of type inflation due to orthographic difference in the corpus occurs in relation to variable use of macrons and unmarked long vowels. Some texts are written using macrons and some without, meaning that the same spoken word will be represented as two different written types in the corpus. Furthermore, there are a high number of errors associated with the use of macrons, which occur both when a writer using macrons has omitted macrons (for example the use of **nga* where the writer typically uses *ngā*), and when writers have used macrons in words which are typically written with short vowels (such as **māhi*). The high frequency of such errors may reflect the finding that younger speakers of te reo Māori differentiate the pronunciation of long and short vowels far less than speakers in the late 1800s-early 1900s (Harlow, 2007, pp. 80-81). In all, the corpus contains 6,774 types with macrons, of which 3,133 correspond to types in the corpus which are identical apart from the macrons. Of course not all of these words are errors. We can categorise these macron/no macron types in three (overlapping) ways:

- Types which represent different morphemes such as *ki* ‘to’ or ‘at’ and *kī* ‘say’ or ‘full’.
- Types which represent the use of different orthographies such as *nga* and *ngā*.
- Types which represent clear errors such as **Māōri*.

These categories overlap because words such as *nga* may be used intentionally in one text where the writer is using unmarked long vowels, but used erroneously in another text where the writer is using macrons to represent long vowels.

We can examine the distribution of the different long vowel orthographies in the corpus by looking at the distribution of the highest-frequency word which contains a long vowel, namely *ngā* (alternatively *nga* or *ngaa*)[3]. Of the 577 texts in the post-1970 corpus, 570 contain at least one instance of *ngā*, *nga* or *ngaa*. Because nearly all texts in the corpus use at least one variant of *ngā, nga* or *ngaa* (the remaining seven texts that do not are very short), we can use it to assess the use of different vowel systems in the contemporary corpus. The distribution of *ngā/nga/ngaa* throughout different texts in the corpus is as follows:

All texts containing *ngā*	412
All texts containing *nga*	343
All texts containing *ngaa*	13
Texts with both *ngā* and *nga*	185
Texts with both *ngā* and *ngaa*	4
Texts with both *nga* and *ngaa*	13

From this distribution we can see that macrons appear to be the most widely-used orthographic system in the corpus. This assumption is supported by the token count: *ngā* occurs 64,769 times in the corpus, approximately five times as often as *nga*, which occurs 12,038 times. *Ngaa* occurs 1,446 times.

What is striking is the high proportion of texts that contain both *ngā* and *nga*. We can ascertain that *nga* is likely to be an error by looking at the how frequently it occurs in a text. In texts that use

[3] In doing this, I am assuming that the distribution of *ngā/nga/ngaa* reflects the typical distribution of long vowel orthography throughout the corpus. The advantage of doing this is that this process captures most of the texts in the corpus in a straightforward way. However, the distribution of *ngā/nga/ngaa* may not be a completely accurate reflection of orthographic systems in the corpus, especially in the case of *ngaa*, where the number of texts involved is small and many of the instances of *ngaa* occur as part of the iwi name *Ngaa Rauru Kitahi*.

macrons *ngā* appears on average 38 times per thousand words. It could be expected that where authors are intentionally using unmarked long vowels, the distribution of *nga* would be similar, but where *nga* appears in error in texts that use macrons, it would appear as a low-frequency term. This assumption is validated by looking at the distribution of *ngā* and *nga* in the corpus, as illustrated in figure 1. The frequencies of each term in the texts were grouped into bands of five words per thousand. The blue series shows that in texts containing *ngā* the term appeared most frequently between 35-40 times per thousand words. Texts that contained very high or low rates of *ngā* were typically shorter texts where a few instances of the term could markedly affect the rate per thousand words. In contrast to the smooth, unimodal distribution of *ngā*, however, the red series representing *nga* spikes sharply in the 0-5 words per thousand band. That a large number of texts contain very few instances of *nga* indicates that this term is likely to slipped in due to the omission of the macron in texts that otherwise use *ngā*.

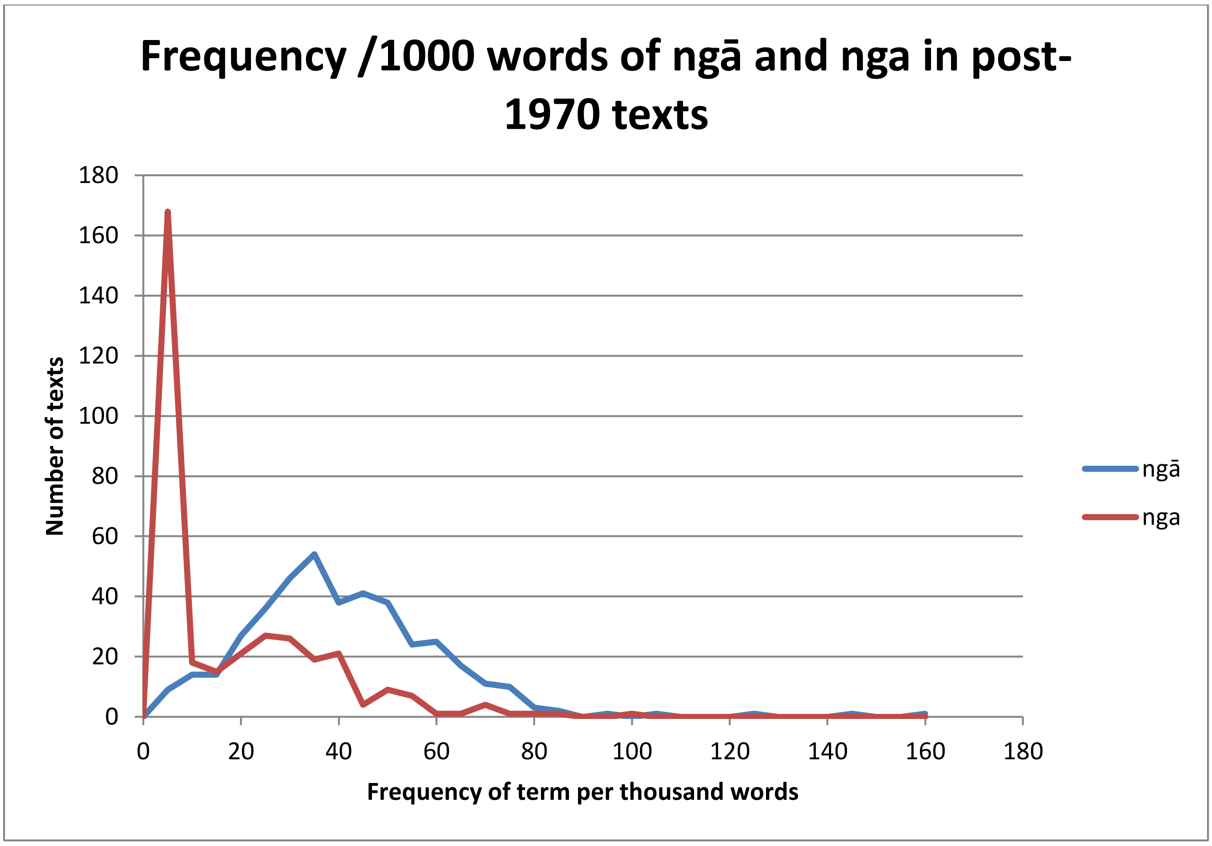

Figure 1 Frequency of ngā and nga per thousand words in post-1970 texts

Scanning the wordlist for macron errors resulted in the identification of 1,361 possible error types (around 20% of the types with macrons and just under 6% of the total number of types). It is likely that a more detailed examination of the wordlist, which would need to include looking at words in context, would yield a greater number of errors.

Double vowels

In order to determine the degree to which double vowel orthography was used in the corpus, I searched for the highest frequency term in the corpus that could be written either with double vowels or macrons, namely *ngā/ngaa.* The search retrieved 13 files out of 577 which contained *ngaa,* of which ten made extensive use of double vowel orthography. One of the remaining files contained only a few words with double vowels and two of the files only used *ngaa* as part of the tribal name *Ngaa Rauru Kitahi*. The texts using double vowels extensively appeared to be clustered around Tainui and and Ngaa Rauru Kitahi, suggesting that double vowels are utilised by only a few isolated writers within the corpus.

I then compiled a wordlist from these files and cross-checked it against the full corpus wordlist for unique types that only occurred in these ten texts. The double vowel texts contained 935 unique types, of which 482 types contained double vowels. These 482 types with double vowels thus make up around 2% of the types in the corpus.

Lexical variation

Dialect variation

Dialect variation does not appear to have been responsible for a great increase in the number of types in the corpus. It appears that regional differences in pronunciation such as the merger of *n and *ŋ to /n/ in the Bay of Plenty, and the South Island merger of *k and *ŋ to /k/ are not typically reflected in the orthography of post-1970 texts in the corpus (although three texts do contain brief examples of South Island dialect reflected in the orthography). Users of the corpus interested in finding texts which reflect characteristics of particular dialects may be able to find texts of interest by looking for alternate forms of high-frequency words associated with a particular dialect, such as the variant forms of determiners such as *tēnei* or *ētahi* (e.g. *tēneki* for *tēnei*, or *wētahi* for *ētahi*).

Alternative morphology

There are a number of base words in the corpus that have a large number of associated passive and nominal forms. This variety, combined with reduplication, means that in some cases one root morpheme may be represented by many words in the corpus. This appears to be particularly true of words that have acquired expanded meanings in modern Māori, especially when these words were not previously used as verbs. For example, for the term *arotake* “review, evaluate, audit” which occurs 926 times in the post-1970 corpus, there are the following passives:

Arotakengia	50 occurrences
Arotakehia	46 occurrences
Arotakea	36 occurrences
Arotaketia	19 occurrences

There are also the following nominalisations:

Arotakenga	216 occurrences
Arotakehanga	11 occurrences
Arotaketanga	3 occurrences

While there is an established but not absolute preference for the nominalisation *arotakenga*, there are a number of accepted forms of the passive with no one term dominating. Future research may be able to identify factors (for example, dialect, date of publication or individual variation) that may affect the distribution of these allomorphic suffixes.

In the case of at least a small number of words, different nominal forms can carry different shades of meaning. *Te Aka Māori-English dictionary* (Moorfield, 2013) distinguishes between *akomanga* "classroom, class", *akoranga* "learning, subject, discipline", and *ākonga* "student, pupil, learner" (similarly, it lists *akonga* without a macron in the phrase *akonga a mua* "alumnus"). It is not clear how many of the terms in the Māori Legal Corpus with alternative suffixes may fall into this category of words with differentiated meanings.

Conclusion

Upwards of 10% of the types in the post-1970 corpus appear to be errors. These errors are a combination of misplaced macrons, most of which are likely to exist in the original texts, and errors that have been introduced in the process of digitising texts. This estimate was made by reviewing word list without checking the context for each word; because of the length of the list and my incomplete knowledge of te reo Māori it is likely that some errors will have been overlooked and other non-error types included mistakenly. The majority of non-macron errors appear to have been introduced either as run-on words or through the substitution of similar-looking characters. A more detailed analysis of these introduced errors using high-frequency words indicated that word boundary errors accounted for around 60% of the errors in the sample and letter substitution accounted for about 40%. Because a conservative approach was taken to categorising run-on words that could possibly have been compound words, it is likely that the true number of introduced run-on errors in higher than this. This could be checked by looking at the original texts to determine whether the author did use a compound word or not, but this would be a very time-consuming process.

Areas for further research

This initial examination of the corpus has focussed on what can be said about errors and orthographic variation in the corpus. It has not examined in any level of detail the reasons for other forms of language diversity such as reo-a-iwi or change in the language through time. These areas of non-error variation in the language may be of interest for further study. In terms of the diversity of types in the corpus, analysis of compound words and transliterations may be particularly fruitful.

The research has examined te reo Māori at the level of individual words. But as Biggs (1973) identifies, the phrase rather than the word is the natural unit of te reo Māori. Although analysis of the corpus at a phrasal level was outside the scope of this project, the software (WordSmith Tools) used to analyse the corpus also offers powerful tools for analysing the language at a phrasal level, and this kind of analysis may be of interest to future researchers.

Another avenue for investigation would be an examination of the number of hepaxlegomena (words that only occur once) as a means of estimating rates of error. This could be done through comparing the number of hepaxlegomena in the Legal Māori Corpus to similar corpora such as the Māori Broadcast Corpus, and also looking at the comparative rate of hepaxlegomena in legal language compared to in general corpora in other languages such as English.

References

Biggs, B. (1973). *Let's learn Maori : a guide to the study of the Maori language*. Wellington, New Zealand: Reed Education.

Boyce, M. (2006). *A corpus of modern spoken Māori.* (Ph.D.), Victoria University of Wellington, Wellington, New Zealand. (1001855, 11177831)

Darwin, J., & Stephens, M. (2009). The Legal Maori Archive: Construction of a large digital collection. *The New Zealand Library & Information Management Journal, 51*(3), 161-171.

Harlow, R. (2007). *Māori: A linguistic introduction*. Cambridge, England: Cambridge.

Kilgarriff, A. (2006). *BNC database and word frequency lists*. Retrieved from: http://www.kilgarriff.co.uk/bnc-readme.html#lemmatised

Moorfield, J. C. (Ed.) (2013) Te aka Māori-English, English-Māori dictionary and index [online version].

Scott, M. (2010). WordSmith Tools (Version 5th).

Te Taura Whiri i te Reo Māori. (2012). *Guidelines for Māori language orthography*. Te Taura Whiri i te Reo Māori. Retrieved from http://www.tetaurawhiri.govt.nz/english/pub_e/downloads/Guidelines_for_Maori_Language_Orthography.pdf