Diacritic Restoration and the Development of a Part-of-Speech Tagset for the Māori Language

A thesis

submitted in partial fulfilment

of the requirements for the degree

of

Master of Science in Computer Science

at

The University of Waikato

by

JOHN COCKS

The University of Waikato

2012

Abstract

This thesis investigates two fundamental problems in natural language processing: diacritic restoration and part-of-speech tagging. Over the past three decades, statistical approaches to diacritic restoration and part-of-speech tagging have grown in interest as a consequence of the increasing availability of manually annotated training data in major languages such as English and French. However, these approaches are not practical for most minority languages, where appropriate training data is either non-existent or not publically available. Furthermore, before developing a part-of-speech tagging system, a suitable tagset is required for that language. In this thesis, we make the following contributions to bridge this gap:

Firstly, we propose a method for diacritic restoration based on naive Bayes classifiers that act at word-level. Classifications are based on a rich set of features, extracted automatically from training data in the form of diacritically marked text. This method requires no additional resources, which makes it language independent. The algorithm was evaluated on one language, namely Māori, and an accuracy exceeding 99% was observed.

Secondly, we present our work on creating one of the necessary resources for the development of a part-of-speech tagging system in Māori, that of a suitable tagset. The tagset described was developed in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora, and was the result of in-depth analysis of the Māori grammar.

Acknowledgements

I would like to extend my gratitude to my supervisor Prof. Te Taka Keegan for his help, guidance and support throughout my Masters degree.

I would also like to thank the University of Waikato and the National Institute of Research Excellence for Māori Development and Advancement, for making my dream of pursuing a Masters degree a reality by giving me financial assistance.

Finally I would like to thank the following people (in no particular order) for their contributions to this research: Hemi Whaanga, David Bainbridge, Stuart Yeates, Kevin Scannell, Ian Witten and Michael William Taiapa.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis investigates two fundamental problems in natural language processing: diacritic restoration and part-of-speech tagging. Diacritic restoration, also known as accent restoration, is the problem of inserting diacritics into a text where they are otherwise missing. The automatic insertion of diacritics into a text is a vital pre-processing step in various natural language processing applications, such as Corpora Acquisition, Information Retrieval and Machine Translation. Part-of-speech tagging, also known as grammatical tagging, is the process of assigning a syntactic category such as a noun, verb, pronoun, adjective or other lexical class marker to each word in a running text. Part-of-speech tagging is required for several natural language processing tasks such as Speech Recognition, Information Extraction and Word Sense Disambiguation.

During the last three decades, statistical approaches to the problems of diacritic restoration and part-of-speech tagging have grown in interest as a consequence of the increasing availability of annotated corpora in major languages such as English and French. However, these approaches are not practical for most minority and resource-scarce languages such as Māori, where appropriate training data is either non-existent or not publically available. Moreover, the process of manually annotating training data is both a time-consuming and expensive task, requiring trained human annotators with substantial amounts of supervision. Therefore, the only viable alternative for minority languages is to employ techniques and approaches different from those which are commonly used in natural language processing (Streiter, 2003).

As previously mentioned, the Māori language is a minority language and is the indigenous language of New Zealand. During the eighteenth century, Māori was the predominant language of New Zealand. However, in 2006 it was estimated that only 4% of New Zealanders could speak Māori. This rapid decline in

speakers was largely contributed to by the influence that the English language had on Māori. Fortunately over the past four decades, major initiates have brought about a revival in the language. Nonetheless, although the Māori language is no longer in an unstable state, natural language processing of Māori is still in its infancy. Therefore, the development of a diacritic restoration algorithm and a part-of-speech tagger is crucial, as they are necessary pre-processing steps to subsequent natural language processing tasks. In this thesis we make the following three contributions to this area of research:

Firstly, we propose a method for diacritic restoration based on naive Bayes classifiers that act at word-level. Classifications are based on a rich set of features, extracted automatically from training data in the form of diacritically marked text. This method requires no additional resources, which makes it language independent. The algorithm was evaluated on one language, namely Māori, and an accuracy exceeding 99% was observed.

Secondly, we describe the Māori Macron Restoration Service, a web-based application developed as part of this thesis for diacritic restoration in Māori. This application enables users to restore diacritics in text via diacritic input or file upload.

Finally, we present our work on creating one of the necessary resources for the development of a part-of-speech tagging system in Māori, that of a suitable tagset. The tagset described was developed in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora, and was the result of in-depth analysis of the Māori grammar.

1.1 Structure of the Thesis

This thesis is structured in a manner which incorporates the relevant literature review, findings and discussions in each chapter where necessary, and is organized as follows:

- Chapter 2 focuses on the problem of diacritic restoration in Māori, incorporating the relevant literature review, findings and discussions.

- Chapter 3 describes the web-based application for diacritic restoration in Māori, based on the algorithm described in Chapter 2.
- Chapter 4 outlines the proposed tagset for part-of-speech tagging in Māori.

Chapter 2

Diacritic Restoration

The Māori language, along with other Polynesian languages, features a written diacritical mark above vowels, signifying a lengthened pronunciation of the vowel. Māori texts without diacritics are quite common in electronic media. The problem arises as most keyboards are designed for English and the process of inserting diacritics becomes laborious. In all but the most ambiguous cases, a native reader can still infer the writer's intended meaning. However, the absence of diacritics can still confuse or slow down a reader and it makes pronunciation and meaning difficult for learners of the language.

For other languages using diacritics, such as German and French, this problem can typically be handled by a simple lexicon lookup procedure that translates words without diacritics into the properly marked format (Wagachar and Pauw, 2006). However, this is not the case for resource-scarce or minority languages such as Māori, where large lexicons are either non-existent or not readily available.

In this thesis, we propose a machine learning approach to diacritic restoration based on a naive Bayes classifier that acts at word-level. The proposed algorithm predicts the placement of diacritics on the basis of local word context. The algorithm is contrasted with a traditional grapheme-level algorithm, originally proposed by Scannell (2010), and shows a significant increase in accuracy for diacritic restoration in Māori text.

The remainder of this chapter is organized as follows. Section 2.1 reviews previous work on diacritic restoration. Section 2.2 discusses diacritics in Māori. Section 2.3 outlines the dataset used for training and testing purposes. Section 2.4 presents the baseline algorithm for diacritic restoration. Sections 2.5 and 2.6 describe our proposed approach to diacritic restoration and the features employed,

respectively. Finally, in section 2.7, we present experimental results for the proposed diacritic restoration algorithms.

2.1 Previous Work

Diacritic restoration, also known as accent restoration, is the problem of inserting diacritics into a text where they are otherwise missing. The automatic insertion of diacritics into a text is a vital pre-processing step in various natural language processing applications, such as Corpora Acquisition, Information Retrieval and Machine Translation.

Up until recently, the majority of research on diacritic restoration has been directed towards major languages such as French and German, and less emphasis directed towards minority languages. The centre of attention is most likely due to the availability or lack thereof of natural language processing resources, such as part-of-speech taggers, which are commonly used in modern diacritic restoration methods.

In recent past, Crandall (2005) proposed an HMM-based method for diacritic restoration that uses a morphological analyzer. The problem with this method is that morphological analyzers are almost non-existent in resource-scarce languages. Furthermore, the development of a morphological analyzer is known to be a time-consuming and expensive task for any given language.

Mihalcea and Nastase (2002), propose a different method for diacritic restoration based on learning mechanisms that act at the grapheme-level. This method was evaluated on four languages, namely Czech, Hungarian, Polish and Romanian, and an average accuracy of over 98% was observed. The advantage with this method is that no additional natural language processing resources or tools other than raw text is required.

Pauw and Wagacha (2006), describe a similar method to the problem of diacritic restoration, but rely on a memory-based learner for classifying instances. Under this method, scores exceeding 90% were reported for numerous African languages, as well as Vietnamese and Chinese Pinyin.

Recently, Scannell (2010), extended upon the work of Pauw and Wagacha (2007) by employing a naive Bayes classifier that also acts at grapheme-level; reporting a high degree of accuracy for numerous languages using training data in the form of a web-crawled corpus. Interestingly, this work reported an accuracy of 97.5% for diacritic restoration in Māori. This score represents an increase of 1% over the reported baseline method which chooses the most frequent diacritic pattern in the training set.

In this thesis, we further extend the work of Scannell (2010) by employing a naive Bayes classifier that acts at word-level opposed to the grapheme-level. In order to determine the feasibility of the proposed approach, the experiments outline by Scannell are reproduced using a large, high-quality corpus of Māori texts, and the results are contrasted with those obtained from the proposed word-level algorithms.

2.2 Diacritics in Māori

The Māori alphabet consists of 15 letters that can be extended to a set of 20 by additional marks and vowels. The 15 letters consist of 10 consonants and 5 vowels: a, e, h, i, k, m, n, ng, o, p, r, t, u, w and wh. Vowels in Māori can be pronounced either short or long, so in written form, long vowels carry a diacritical mark. A diacritical mark, also known as a macron in Māori, is a short stroke placed above the lengthened vowel. Table 1 shows the complete set of short and long vowels in Māori.

Short vowel	**Long vowel**
a	ā
e	ē
i	ī
o	ō
u	ū

Table 1: Short and long vowels in Māori

A Māori text without diacritics will substitute long vowels for short vowels. Consequently, this causes considerable ambiguity at the level of the word, as many words with different vowel patterns occur in identical diacritic-less settings.

The word *ana*, for example, has 4 possible forms that have valid interpretations when diacritized. It may have the interpretation of the noun *cave* in *ana*, or interpreted as *yes* in the interjection form *āna*. It can also be interpreted as *there* in the locative place form *anā*, or interpreted as a determiner in the form *ānā*.

2.3 Dataset

The diacritic restoration algorithms presented in this thesis were trained and evaluated on a large corpus of Māori text with near perfect diacritization. The corpus consists of old Māori scripts, short stories, bible verses, dictionary definitions and conversational texts. This corpus was diacritized at the University of Waikato by Māori language specialists in the faculties of Māori and Pacific Development, and Computing and Mathematical Sciences. Table 2 displays statistical data extracted from the corpus.

	Total
Words	3,739,139
Words with diacritics	860,038 (23.0%)
Distinct words	11,996
Characters	13,088,331
Characters with diacritics	905,467 (6.92%)
Distinct characters	34

Table 2: Statistical data extracted from the diacritic corpus

As seen in the table above, the corpus contains a total of 3,739,139 words, of which approximately 12 thousand are distinct. Furthermore, roughly a quarter of the words in the corpus contain at least one diacritical letter. That is, on average, one diacritical character occurring in every 14 characters. Note: a word is defined here to be any sequence of alphanumerical characters delimited by a whitespace character or a punctuation mark; whereas a character is defined here to be any character excluding whitespace characters, punctuation marks and symbols.

In order to gauge the difficulty of the diacritic restoration problem, we further analyze the corpus as described in a previously published paper by Tufis (2011).

In this paper, Tufis categorizes words under two main categories: U-words and A-words. We extend this categorization by including a third category: O-words. A full description of these categories follows.

Word Category	**Total**
U-words	467,079 (12.49%)
A-words	1,690,791 (45.22%)
O-words	1,581,269 (42.29%)

Table 3: U-word, A-word and O-word categorization of words

The first category, U-words, includes those words which are unambiguous and if missing diacritics are not legal words of Māori. A word is considered unambiguous if it can only stand for a single word in a diacritic-less setting. Examples of U-words in Māori are: *hauku* (haukū – damp), *mawe* (māwe – talisman), *mera* (mēra – mail), *ngawe* (ngawē – howl) and *pukei* (pūkei – heap). Note, the Māori word and English translation are enclosed in parentheses. U-words account for roughly 12% of the words in the corpus, as shown in Table 3.

The second category, A-words, includes those words which are ambiguous and if missing diacritics could stand for one of multiple words. For example, in a text where diacritics have been omitted, the word *tete* could stand for any of the following words: tete – javelin; tētē – teal. A-words account for approximately 45% of the words in the corpus.

The third and final category, O-words, includes those words which are unambiguous and if missing diacritics are legal words (i.e. words which do not have any diacritical letters in a diacritized setting). Examples of O-words are: ahao – spear; huku – tail; maheni – magazine; ngariri – love; whakatio – freezer. U-words account for 42% of the words in the corpus.

Fortunately, both U-words and O-words can be disambiguated using a simple dictionary lookup procedure that checks the lexicon for unambiguous words and disambiguates them accordingly. This simple procedure accounts for

approximately 55% of the words in the corpus. Of the remaining 45%, some form of disambiguation is required.

2.4 Baseline Algorithms

In order to determine the significance of the proposed diacritic restoration algorithms, two baseline algorithms are adopted, which were initially defined by Crandall (2005). The first baseline algorithm assumes no diacritical markings exist. That is, it simply outputs the diacritic-less word that it receives as input. This baseline algorithm gives an accuracy of 79.94% for Māori, or an error once in every five words. These results are discussed further in Section 2.7.

The second baseline algorithm which was adopted here, identifies all possible diacritic patterns for a given diacritic-less word, and chooses the most dominant pattern observed. In cases where multiple patterns are observed equally often, the algorithm chooses one pattern at random. This baseline algorithm achieves a mean accuracy of 97.11% for Māori, which is significantly higher than the accuracy of 79.94% reported for the first baseline algorithm. The high accuracy shows that in many cases where there is an ambiguity in Māori, that one pattern is generally more dominant. Thus a high degree of accuracy can be achieved by selecting the most common diacritical pattern for each word. Table 4 shows examples of diacritic pattern distributions extracted for Māori.

Diacritic-less Word	**Diacritic Pattern**	**Number**	**%**
ana	ana	2114	16
ana	āna	5548	42
ana	anā	2642	20
ana	ānā	2906	22
tipa	tipa	2440	76
tipa	tīpā	770	24
popo	popo	6476	76
popo	pōpō	1789	21
popo	pōpō	256	3

Table 4: Diacritic pattern distributions of Māori words

2.5 Learning Algorithms

We formulate the task of restoring diacritics as a classification problem, where a label (i.e. diacritical pattern) is assigned to each grapheme or word in a diacritic-less text. For the purposes of our experiments, we decided to use a naive Bayes classifier. The reasons for this are twofold. Firstly, in spite of their naive design, naive Bayes classifiers are widely used in various classification tasks in natural language processing. Secondly, naive Bayes classifiers are efficient in terms of training and testing times (Mihalcea, 2002). What follows is a formal definition of the naive Bayes classifier and its application to diacritic restoration.

Naive Bayes classifiers are a set of probabilistic learning algorithms based on applying Bayes' theorem with the naive assumption of independence between features. Given a class variable c and a dependent feature vector x_i through x_n, Bayes' theorem states the following relation:

$$P(c \mid x_1, x_2, ..., x_n) \propto P(c) \prod_{i=1}^{n} P(x_i \mid c) \tag{1}$$

Where $P(c)$ is interpreted as the conditional probability of class c occurring, and $P(x_i \mid c)$ is interpreted as the conditional probability of attribute x_i occurring given class c. In order to find the most likely classification $\hat{c}$, given the attribute values x_1 through x_n, equation (1) can be rewritten as:

$$\hat{c} = \arg\max P(c) \prod_{i=1}^{n} P(x_i \mid c) \tag{2}$$

In practice, equation (2) often results in a floating point underflow as n increases. It is therefore better to perform the computation by adding logarithms of probabilities instead of multiplying probabilities as in (3).

$$\hat{c} = \arg\max \left[\log P(c) + \sum_{i=1}^{n} \log P(x_i \mid c) \right] \tag{3}$$

In order to apply a Naive Bayes classifier to the task of restoring diacritics, estimates for the parameters $P(c)$ and $P(x_i \mid c)$ in equation (3) are required. Since estimating the parameters for the grapheme-level and word-level approaches are similar, we will only demonstrate one of them, namely the word-level approach.

Assuming a diacritically marked text T is a sequence of words w_i through w_n, where n is the total number of words in the text, then T can be represented as:

$$T = w_1, w_2, \ldots, w_n \tag{4}$$

Further, assume each word w_i in T has an associated word form b_i, where b_i is the result of removing all diacritics from w_i. Thus, a text T has a word form sequence T_b associated with it, which can be written as follows:

$$T_b = b_1, b_2, \ldots, b_n \tag{5}$$

Now let W_d be the set of distinct words in T and let B_d be the set of distinct word forms in T_b. Further, let $f : B \rightarrow W_s$ be a function that maps a word form b_i to a set of words W_s, where $W_s \subseteq W_d$, and each word in W_s has a corresponding word form equal to b_i. The goal is to find, for each word form b_i in T_b, the word w in $f(b)$, such that w maximizes the probability for all words in $f(b)$. Using Bayes theorem in (3), the prior probability for each word w in $f(b)$ can be estimated by:

$$P(w) = \frac{N_w}{N} \tag{6}$$

Where N_w is the total number of times word w occurs in text T, and N is the total number of times each word in $f(b)$ occurs in text T. Further, the conditional probability for each word w in $f(b)$ is estimated as:

$$P(w) = \frac{N_{w_i} + 1}{N_i + n} \tag{7}$$

Where N_{w_i} is the total number of times word w with feature i occurs in text T, and N_i is the total number of times each word w in $f(b)$ with feature i occurs in text T, and n is the total number of words in $f(b)$. To avoid zero estimates, Laplace smoothing is employed.

2.6 Features

Within the naive Bayes framework, any type of features can be used. However, as previously mentioned, there are currently few natural language processing tools available to facilitate the extraction of features in Māori, which are common in other machine learning diacritic restoration algorithms. As a result, the features employed here require no particular processing other then tokenization. These features are divided into two categories: grapheme-level and word-level features. Both features are discussed in the following subsections.

2.6.1 Grapheme-level Features

Scannell (2010) employs a naive Bayes classifier that acts at the level of the grapheme, reporting a high degree of accuracy for numerous languages. These classifiers are trained using various sets of features, each consisting of n-grams of consecutive graphemes relative to the target grapheme. Each n-gram is represented by the vector (o, n), where o represents the offset of the n-gram from the target grapheme, and n represents the length of the n-gram. These feature sets are outlined below in Table 5. Note: in this thesis, we propose a new grapheme-level feature set: FSG5.

Name	**Features**	**Description**
FSG1	(-3, 1), (-2, 1), (-1, 1), (1, 1), (2, 1), (3, 1)	The three graphemes on each side of the target grapheme.
FSG2	(-5, 1), (-4, 1), (-3, 1), (-2, 1), (-1, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1)	The five graphemes on each side of the target grapheme.
FSG3	(-4, 3), (-3, 3), (-2, 3), (-1, 3), (0, 3), (1, 3), (2, 3)	The two trigrams on each side of the target grapheme, and the three

		trigrams overlapping the target grapheme.
FSG4	(-3, 3), (-1, 3), (1, 3)	The trigram on each side of the target grapheme, and the trigram centered on the target grapheme.
FSG5	(-2, 5), (-3, 5), (-1, 5)	The 5-gram centered on the target grapheme and the 5-grams starting at offsets -3 and -1.

Table 5: Grapheme-level features for diacritic restoration

2.6.2 Word-level Features

In this thesis, we improve upon previously mentioned approaches to diacritic restoration for Māori by employing a naive Bayes classifier that acts at the level of the word opposed to the grapheme. These feature sets are outlined below in Table 6.

Name	**Features**	**Description**
FSW1	(-1, 1)	The word preceding the target word.
FSW2	(-2, 2)	The bigram preceding the target word.
FSW3	(-3, 3)	The trigram preceding the target word.
FSW4	(1, 1)	The word following the target word.
FSW5	(1, 2)	The bigram following the target word.
FSW6	(1, 3)	The trigram following the target word.
FSW7	(-1, 1), (-2, 2)	The word and bigram preceding the target word.
FSW8	(1, 1), (1, 2)	The word and bigram following the target word.
FSW9	(-1, 1), (1, 1)	The word on each side of the target word.
FSW10	(-2, 2), (-1, 1), (1, 1), (1, 2)	The word and bigram on each side of the target word.
FSW11	(-1, 3), (-2, 2), (1, 2), (-1, 4), (-	The trigram centered on the target

	2, 4)	word, and the bigrams on each side of the target word, and the 4-grams starting at offsets -1 and -2.

Table 6: Word-level features for diacritic restoration

2.7 Results

In order to test the accuracy and robustness of the diacritic restoration algorithms, a ten-fold cross-validation methodology was employed. Under this method, the dataset described in Section 2.3 is randomly partitioned into 10 equal-sized subsets. Of the 10 subsets, a single subset is retained as the validation data for testing purposes, while the remaining 9 subsets are used as training data. During testing, the validation data is artificially stripped of diacritics prior to applying a diacritic restoration algorithm. The cross-validation process is repeated a total of 10 times, with each subset used exactly once as the validation data. The algorithms accuracy is the average of the ten-fold cross validation runs and is reported in terms of the proportion of correctly diacritized words.

The experimental results shown in Table 7 indicate that the word-level naive Bayes algorithms significantly outperform the grapheme-level naive Bayes algorithms, for diacritic restoration in Māori. Note that the baseline algorithms, that of baseline-1 and baseline-2, are not feature sets but are included here for completeness and for comparison purposes.

Feature Set	**Accuracy (%) (proportion of words)**
Baseline-1	79.94
Baseline-2	97.11
FSG1	79.94
FSG2	79.94
FSG3	84.45
FSG4	87.02
FSG5	95.07
FSW1	98.50

FSW2	98.33
FSW3	97.94
FSW4	98.28
FSW5	98.34
FSW6	98.01
FSW7	98.65
FSW8	98.54
FSW9	98.65
FSW10	98.85
FSW11	99.01

Table 7: Results of the baseline, grapheme-level and word-level algorithms

Evidently, the FSW11 feature set resulted in the highest accuracy of 99.01%. This result represents an increase of 1.9% over the baseline-2 algorithm which chooses the most frequent pattern in the training data. As previously mentioned, the FSW11 feature set contains five features: the trigram centered on the target word; the bigram on each side of the target word; the n-grams of length 4 starting at offsets -1 and -2.

A paired t-test was performed to determine if the increase in accuracy between the FSW11 feature set and the baseline-2 algorithm was statistically significant. The mean increase in accuracy (M=1.8928, SD=0.0234, N=10) was significantly greater than zero, t(9)=255.68, two-tail p=1.08989E-18, providing evidence that the FSW11 feature set has a significant increase in accuracy over the baseline-2 algorithm. A 95% C.I. about mean accuracy increase is (1.8761, 1.9096).

Chapter 3

Web-based application for diacritic restoration

This chapter describes the Māori Macron Restoration Service, a web-based application for diacritic restoration in Māori text. The first section of this chapter presents a brief overview of the system. The following sections describe the application architecture in more detail, illustrating the user interface and exploring its features and functions.

3.1 Overview

We have developed a web-based application based on the diacritic restoration algorithms described in the previous chapter. The web-based application, known as the Māori Macron Restoration Service, allows users to automatically restore diacritics in Māori text via direct input or file upload. The application is located on the Greenstone server within the Faculty of Computing and Mathematical Sciences at the University of Waikato (available at http://www.greenstone.org/macroniser as of Dec, 2011).

3.2 The interface

The Māori Macron Restoration Service web-based application is composed of the following windows:

- Direct input window
- File upload window
- File download window
- About window
- Feedback window

Each of these windows will be described in detail in the following sub sections.

3.2.1 The Direct Input window

Figure 1 is a screenshot of the Direct Input window, which enables users to automatically restore diacritics in Māori text via direct input. This is the first window that is displayed when opening the web-based application. The window consists of three main components: a large text area [1] for direct input; a submit button [2], which sends the contents of the text area to the server for processing; a checkbox [3] that determines whether or not pre-existing diacritics are preserved.

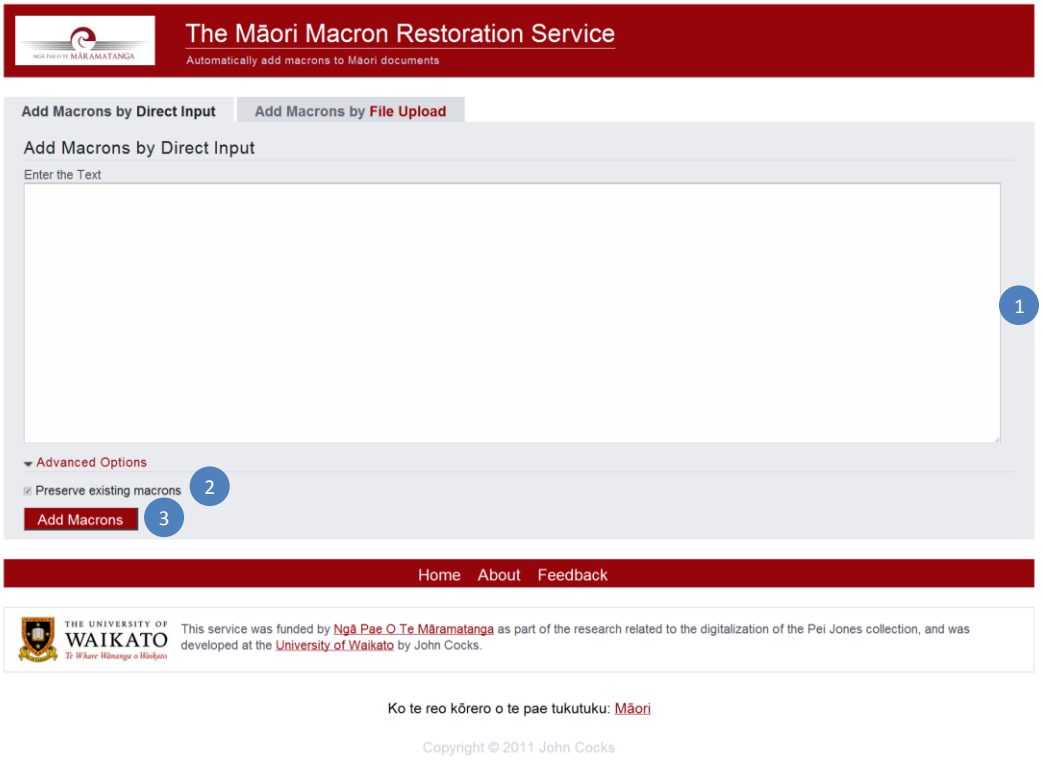

Figure 1: Screenshot of the Direct Input window of the web-based application

3.2.2 The File Upload window

Figure 2 is a screenshot of the File Upload window, which enables users to automatically restore diacritics in Māori text via file upload. This window is accessed through the navigation bar [4] and consists of six main components: a choose file button [5] that displays a dialog box from which the user can select a file from the client's file system; an upload button [9] which sends the specified file to the server for processing; a toggle button [6] that reveals or hides a set of advanced options (discussed below).

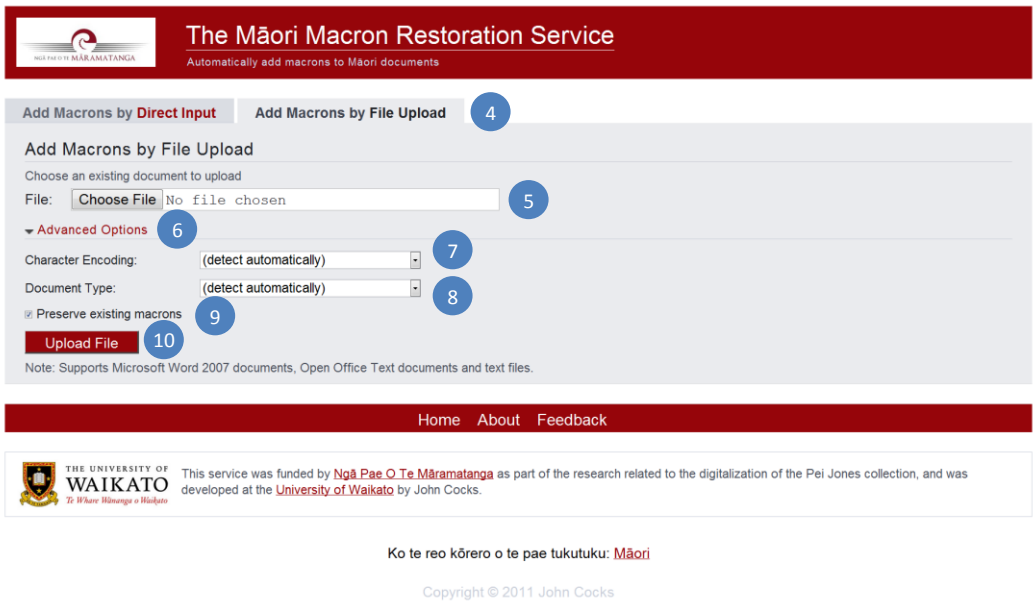

Figure 2: Screenshot of the File Upload window of the web-based application

The set of advanced options include the ability to specific the character set encoding [7] and document type [8] of the file to be uploaded to the server for processing. A checkbox [3] is also provided, which determines whether or not pre-existing diacritics are preserved.

3.2.3 The File Download window

Figure 3 is a screenshot of the File Download window, which enables users to download processed files. The user is redirected to this window after a particular file has been uploaded to the server and successfully processed. This window contains two main components: a label [11] that specifies the uploaded files filename, character encoding and document type; a download button [12] that displays a dialog box from which the user can select a location on the client's file system to save the processed file.

Figure 3: Screenshot of the File Download window of the web-based application

3.2.4 The About window

Figure 4 is a screenshot of the About window. This window is accessed through the navigation bar [13]. The purpose of the About window is to provide an overview of the web-based application, and to give a brief description of the role diacritics play in Māori.

Figure 4: Screenshot of the About window of the web-based application

3.2.5 The Feedback window

Figure 5 is a screenshot of the Feedback window that enables users to anonymously provide feedback, or to report any errors or technical issues related to the web-based application. This window is accessed through the navigation bar [17], and consists of three main components: a text field [14] used to describe the subject matter; a text area [15] to enter the subject matter; a submit button [16], which packages the subject and subject matter into an email and sends it to the administrators of the web-based application.

Figure 5: Screenshot of the Feedback window of the web-based application

Chapter 4

Part-of-speech Tagset

While part-of-speech tagging is an established technology for major languages such as English and French, an array of problems arise while extending these techniques to minority languages such as Māori. One such problem is the process of creating one of the necessary resources for the development of a part-of-speech tagging system, that of a suitable tagset.

For European and East Asian languages, tagsets have matured from mere lists of important morphosyntactic features into hierarchical tagsets, decomposable tags and common frameworks (Baskaran, 2008). To our knowledge, however, no published work exists in the area of tagset design or creation for the Māori language, let alone any other Polynesian language.

In this chapter, we present our work on the development of a part-of-speech tagset for the Māori language. The tagset described was developed in accordance with the EAGLES guidelines for morphosyntactic annotation of corpora (Leech and Wilson, 1996), and was the result of in-depth analysis of the Māori grammar.

The remainder of this chapter is organized as follows. Section 4.1 outlines the EAGLES guidelines for morphosyntactic annotation of corpora. Section 4.2 discusses how the guidelines can be extended to the tagset design of the Māori language. The remaining sections, Sections 4.3 to 4.15, outline each obligatory major word class proposed by the EAGLES guidelines with respect to Māori. Finally, in Section 4.16, the proposed intermediate Māori tagset is defined.

4.1 EAGLES Guidelines

As previously mentioned, the Māori tagset described in this thesis was developed in accordance with the EAGLES guidelines on morphosyntactic annotation (Leech and Wilson, 1996). These guidelines were originally designed to help

standardise tagsets for European and East Asian languages. In recent past, however, several published studies have shown that the EAGLES guidelines can be successfully applied to other languages. Such was the case for Urdu (Sajjad and Schmid, 2009) and Arabic (Alqrainy and Ayesh, 2005).

The intent of the EAGLES guidelines is to promote standardisation, interchangeability and reusability of annotated corpora, while discouraging the reinvention of the wheel. As stated in the EAGLES guidelines, it is important to avoid a free-for-all in tagging practises (Leech and Wilson, 1996). The EAGLES guidelines accomplish these objectives by providing a flexible framework that in theory can accommodate all levels of mark-up, without restricting the freedom of the tagset designer.

The framework is based on three levels: obligatory major word classes, recommended attributes and optional attributes. The major word classes include: noun, verb, adjective, pronoun/determiner, article, adverb, adposition, conjunction, numeral, interjection, unique/unassigned, residual, punctuation. The recommended attributes include: person, gender, number, finiteness, tense, void, status, etc. The optional attributes include: countability, aspect, separability, re-exivity, auxiliary, etc. Note, the recommended and optional attributes are organized by the obligatory major word classes, and do not necessarily correspond across word classes.

4.2 Extending the EAGLES Guidelines to Māori

In order to define the linguistic categories of an EAGLES compliant tagset, it is necessary to have a model of the language to categorise (Hardie, 2003). For languages such as Māori, where there has been very little research in tagset design, the only viable option is to derive this model from published descriptions of the grammar. Therefore, we decided to rely on the grammars defined by Harlow (2001) and Bauer (1997), in order to furnish a model of the Māori language.

4.3 Nouns

The EAGLES guidelines propose six attributes for nouns as shown in Table 8. The attributes (i)-(iv) are recommended while the attributes (v) and (vi) are optional.

Attribute	**Values**			
(i) Type	1. Common	2. Proper		
(ii) Gender	1. Masculine	2. Feminine	3. Neuter	4. Common
(iii) Number	1. Singular	2. Plural		
(iv) Case	1. Nominative	2. Genitive	3. Dative	4. Accusative
	5. Vocative	6. Vocative	7. Indeclinable	
(v) Countability	1. Countable	2. Mass		
(vi) Definiteness	1. Definite	2. Indefinite	3. Unmarked	

Table 8: EAGLES guidelines for nouns

Type

With respect to the attribute type (i), Māori nouns can be categorized into three subcategories: common nouns which are virtually always preceded by a determiner when head of a noun phrase; locative nouns which when preceded by a particle like *i* or *ki*, may not have a determiner between; personal nouns which when preceded by a preposition such as *i* or *ki*, must also be preceded by the personal article *a* (Harlow, 2001). For reasons of simplicity, the attribute type (i) is left unchanged, with the exception of the following additions found in Table 9.

(i) Type	3. Locative	4. Personal		

Table 9: Additional values for the attribute type

Number

The attribute number (iii) is not relevant to Māori nouns. Instead the distinction between singular and plural is indicated by the determiner associated with the noun, with one exception. There are a very few words, all of them terms for

people which do have different forms for singular and plural (Harlow, 2001). These are listed in Table 10 and are taken into account in the intermediate tagset.

Singular	**Plural**
wahine	wāhine
tangata	tāngata
matua	mātua
tuahine	tuāhine
tuakana	tuākana
teina	tēina
tipuna	tīpuna
tamaiti	tamariki
whaea	whāea

Table 10: Singular and plural forms for common nouns (Harlow, 2001)

Wh-type

In Māori, wh-type interrogatives exist in the majority of parts of speech. Nouns are no exception. Accordingly, a new attribute is added to the intermediate tagset, as shown in Table 11.

(vii) Wh-type	1. Interrogative				

Table 11: An additional attribute wh-type for nouns

EAGLES attributes for nouns not used in this tagset

The attributes, gender (ii), case (iv), countability (v) and definiteness (vi) are not relevant to Māori nouns. Furthermore, the standard view is that the attribute definiteness (vi) is marked by the determiner associated with the noun (Bauer, 1997). Accordingly, these attributes are ignored in the intermediate tagset.

Intermediate tagset for nouns

Table 12 gives an overview of the intermediate tagset for nouns in Māori.

Description	**Intermediate Tag**	**Example**
Interrogative common	N1000001	aha

noun		
Interrogative locative noun	N3000001	hea
Interrogative personal noun	N4000001	wai
Singular common noun	N1010000	wahine
Plural common noun	N1020000	wāhine
Common noun	N1000000	hātarei
Locative noun	N3000000	konei
Personal noun	N4000000	ākuhata
Noun	N0000000	

Table 12: Intermediate tagset for Māori nouns

4.4 Verbs

For verbs, the EAGLES guidelines propose the attributes (i)-(xiii) of Table 13. The attributes (i)-(viii) are recommended, while the attributes (ix)-(xiii) are optional.

Attribute	**Values**			
(i) Person	1. First	2. Second	3. Third	
(ii) Gender	1. Masculine	2. Feminine	3. Neuter	
(iii) Number	1. Singular	2. Plural		
(iv) Finiteness	1. Finite	2. Non-finite		
(v) Verb form / Mood	1. Indicative	2. Subjunctive	3. Imperative	4. Conditional
	5. Infinitive	6. Participle	7. Gerund	8. Supine
	9. –Ing form			
(vi) Tense	1. Present	2. Imperfect	3. Future	4. Past
(vii) Voice	1. Active	2. Passive		
(viii) Status	1. Main	2. Auxiliary	3. Semi-auxiliary	
(ix) Aspect	1. Perfective	2.		

		Imperfective		
(x) Separability	1. Non-separable	2. Separable		
(xi) Reflexivity	1. Reflexive	2. Non-reflexive		
(xii) Auxiliary	1. Have	2. Be		
(xiii) Aux-function	1. Primary	2. Modal		

Table 13: EAGLES guidelines for verbs

Type

Māori verbs can be categorised into four subcategories: transitive, intransitive, experience and neuter verbs (Harlow, 2007). Arguably, adjective is also a subcategory of Māori verbs. Nonetheless, the EAGLES guidelines regard adjectives as a major category; therefore adjectives are treated independently from verbs. Accordingly, a new attribute type (xiv) is added to the intermediate tagset concerning the subcategories of verb above, as shown in Table 14.

(xiv) Type	1. Transitive	2. Intransitive	3. Experience	4. Neuter

Table 14: Additional attribute type for verbs

Wh-type

The attribute wh-type (xv) is relevant to Māori verbs and is added to the intermediate tagset, as seen in Table 15.

(xv) Wh-type	1. Interrogative			

Table 15: Additional attributes type and wh-type for verbs

EAGLES attributes for verbs not used in this tagset

The EAGLES guidelines recommend a number of attributes that are not relevant to Māori verbs. This is largely due to the occurrence of particles associated with Māori verbs, whose meaning range over tense, aspect and mood (Harlow, 2007). Consequently, the attributes (i)-(xiii) are not considered relevant to Māori verbs. According, these attributes are ignored in the intermediate tagset.

Intermediate tagset for verbs

Table 16 gives an overview of the intermediate tagset for verbs in Māori.

Description	Intermediate Tag	Example
Transitive verb	V000000000000010	patu
Intransitive verb	V000000000000020	oma
Experience verb	V000000000000030	rongo
Neuter verb	V000000000000040	pakaru
Interrogative verb	V000000000000001	aha
Verb	V000000000000000	patu

Table 16: Intermediate tagset for Māori verbs

4.5 Adjectives

The EAGLES guidelines propose seven attributes for adjectives as shown in Table 17. The recommended attributes are (i)-(iv), while the optional attributes are (v)-(vii).

Attribute	Values			
(i) Degree	1. Positive	2. Comparative	3. Superlative	
(ii) Gender	1. Masculine	2. Feminine	3. Neuter	
(iii) Number	1. Singular	2. Plural		
(iv) Case	1. Nominative	2. Genitive	3. Dative	4. Accusative
	5. Vocative	6. Indeclinable		
(v) Inflection-type	1. Weak-Flection	2. Strong-Flection	3. Mixed	
(vi) Use	1. Attributive	2. Predicative		
(vii) NP Function	1. Premodifying	2. Postmodifying	3. Head-function	

Table 17: EAGLES guidelines for adjectives

Wh-type

The attribute wh-type (viii) is relevant to Māori adjectives. Therefore a new attribute is added to the intermediate tagset, as seen in Table 18.

(viii) Wh-type	1. Interrogative	2. Relative	3. Exclamatory		

Table 18: An additional attribute wh-type for adjectives

EAGLES attributes for adjectives not used in this tagset

The attributes (i)-(vii) are ignored in the intermediate tagset because Māori adjectives do not inflect with respect to degree, gender, number or case.

Intermediate tagset for adjectives

Table 19 gives an overview of the intermediate tagset for adjectives in Māori.

Description	Intermediate Tag	Example
Interrogative adjective	AJ00000001	pēhea
Adjective	AJ00000000	pai

Table 19: Intermediate tagset for Māori adjectives

4.6 Pronouns and determiners

For pronouns and determiners, the EAGLES guidelines propose the attributes (i)-(xii) of Table 20. The attributes (i)-(viii) are recommended, while the attributes (ix)-(xii) are optional.

Attribute	Values			
(i) Person	1. First	2. Second	3. Third	
(ii) Gender	1. Masculine	2. Feminine	3. Neuter	4. Common
(iii) Number	1. Singular	2. Plural		
(iv) Possessive	1. Singular	2. Plural		
(v) Case	1. Nominative	2. Genitive	3. Dative	4. Accusative
	5. Non-genitive	6. Oblique	7. Prepositional	
(vi) Category	1. Pronoun	2. Determiner	3. Both	
(vii) Pron.-Type	1. Demonstrative	2. Indefinite	3. Possessive	4. Int./Rel.
	5. Pers./Refl.			

(viii) Det.-Type	1. Demonstrative	2. Indefinite	3. Possessive	4. Int./Rel.
	5. Partitive			
(ix) Special Pronoun Type	1. Personal	2. Reflexive	3. Reciprocal	
(x) Wh-Type	1. Interrogative	2. Relative	3. Exclamatory	
(xi) Politeness	1. Polite	2. Familiar		
(xii) Strength	1. Weak	2. Strong		

Table 20: EAGLES guidelines for pronouns and determiners

Evidently, the EAGLES guidelines treat pronouns and determiners as a single category, due largely to a heavy overlap in their formal and functional characteristics. Moreover, the guidelines recognize that for some descriptions it may be thought best to treat them as different parts of speech (Leech, 1996). As a consequence, the guidelines do not prevent a realignment of categories, but do propose that articles are recognized as a separate part of speech, whether or not included within determiners. Nonetheless, the EAGLES guidelines provide type attributes for pronouns and determiners, pron.-type (vii) and det.-type (viii), of which the intermediate tagset makes use.

Person

With respect to the attribute person (i), Māori like English, makes a distinction between first, second and third person pronouns. However, unlike in English, Māori expresses with more precision quite who the people are included with the speaker in the first person (Harlow, 2001). These are the first person inclusive and exclusive pronouns. Accordingly, new values are added to the attribute person (i) as seen in Table 21.

(i) Person	4. First inclusive	5. First exclusive		

Table 21: Additional values for the attribute person

Number

For the attribute number (iii), Māori makes a distinction between singular and plural pronouns. However, unlike in English, Māori has a set of pronouns which refer to two persons, the dual pronouns. Accordingly, a new value is added to the attribute number (iii) as seen in Table 22.

(iii) Number	3. Dual			

Table 22: Additional values for the attribute number

Pron.-Type

The attribute pron.-type (vii) is not relevant to Māori pronouns. In general, Māori does not have special reflexive or indefinite pronouns (Bauer, 1997). Furthermore, the standard view is that the ordinary pronouns can be used to express reflexivity and indefiniteness; with the result that sometimes a sentence is ambiguous. Of the remaining three values, none are relevant to Māori pronouns, thus are superfluous in the intermediate tagset.

Det.-Type

Māori determiners can be categorized into four main categories: articles, possessive determiners, demonstratives and interrogatives. Note that the EAGLES guidelines propose that articles are recognized as a separate part of speech, of which the intermediate tagset makes use. Nevertheless, the value partitive is not considered relevant to Māori determiners. Therefore, the value is excluded from the intermediate tagset.

Pos.-Form

Of the four categories of Māori determiners outlined above, possessive determiners can be further subcategorized into three subcategories: a-form, o-form and neutral form determiners. These three categories indicate the relationship between the possessor and possessee, which is not a property of either the possessor or possesee (Bauer, 1997). Accordingly, a new attribute is added to the intermediate tagset as shown in Table 23.

(xiii) Pos.-Form	1. A	2. O	3. Neutral	

Table 23: An additional attribute pos.-form for pronouns and determiners

Wh-type

The attribute wh-type (viii) is relevant to Māori determiners.

EAGLES attributes for pronouns and determiners not used in this tagset

The attributes (ii), (iv), (v), (ix), (x) and (xi) are ignored in the intermediate tagset because Māori pronouns and determiners do not distinguish gender, possessive count, case, special pronoun type, wh-type or politeness.

Intermediate tagset for pronouns and determiners

Table 24 gives an overview of the intermediate tagset for pronouns in Māori.

Description	Intermediate Tag	Example
First person inclusive dual personal pronoun	PD4030015000000	tāua
First person inclusive plural personal pronoun	PD4020015000000	tātou
First person exclusive singular personal pronoun	PD5010015000000	au
First person exclusive dual personal pronoun	PD5030015000000	māua
First person exclusive plural personal pronoun	PD5020015000000	mātou
Second person singular personal pronoun	PD2010015000000	koe
Second person dual personal pronoun	PD2030015000000	kōrua
Second person plural personal pronoun	PD2020015000000	koutou
Third person singular personal pronoun	PD3010015000000	ia
Third person dual personal pronoun	PD3030015000000	rāua

Third person plural personal pronoun	PD3020015000000	rātou

Table 24: Intermediate tagset for Māori pronouns

Table 25 gives an overview of the intermediate tagset for Māori determiners.

Description	Intermediate Tag	Example
Demonstrative singular determiner	PD0010020100000	tēnā
Demonstrative plural determiner	PD0020020100000	ēnā
Interrogative singular determiner	PD0010020400000	tēhea
Interrogative plural determiner	PD0020020400000	ēhea
Singular determiner	PD0010020000000	taua
Plural determiner	PD0020020000000	aua

Table 25: Intermediate tagset for Māori determiners

Table 26 gives an overview of the intermediate tagset for Māori possessive determiners.

Description	Intermediate Tag	Example
Possessive singular a-form determiner	PD0010020300001	tāku
Possessive plural a-form determiner	PD0020020300001	āku
Possessive singular o-form determiner	PD0010020300002	tōku
Possessive plural o-form determiner	PD0020020300002	ōku
Possessive singular neutral-form determiner	PD0010020300003	taku
Possessive plural neutral-form determiner	PD0020020300003	aku

Table 26: Intermediate tagset for Māori possessive determiners

4.7 Articles

The EAGLES guidelines recommend the attributes (i)-(iv) of Table 27 for articles. Note that articles are classified under the class of determiners in Māori. Nonetheless, the EAGLES guidelines recommend an individual class for articles. Therefore articles are treated independently from determiners.

Attribute	**Values**			
(i) Article-Type	1. Definite	2. Indefinite	3. Partitive	
(ii) Gender	1. Masculine	2. Feminine	3. Neuter	4. Common
(iii) Number	1. Singular	2. Plural		
(iv) Case	1. Nominative	2. Genitive	3. Dative	4. Accusative
	5. Vocative	6. Indeclinable		

Table 27: EAGLES guidelines for articles

Article-Type

In terms of the attribute article-type (i), there are three types of articles in Māori: the definite articles *te* and *ngā*; the indefinite article *hē;* the personal article *a*. Furthermore, there is no counterpart in English for the personal article *a* (Harlow, 2001). Therefore, for reasons of simplicity, the personal article is not included here; rather it is included in the unique/unassigned category.

Number

With respect to the attribute number (iii), Māori distinguishes between singular and plural for the definite articles. Thus, the attribute is considered relevant and added to the intermediate tagset.

EAGLES attributes for articles not used in this tagset

The attributes gender (ii) and case (iv) are not relevant to Māori articles. Accordingly, these attributes are ignored in the intermediate tagset.

Intermediate tagset for articles

Table 28 presents the intermediate tagset for articles.

Description	Intermediate Tag	Example
Definite singular article	AT1010	te
Definite plural article	AT1020	ngā
Indefinite article	AT2000	hē

Table 28: Intermediate tagset for Māori articles

4.8 Adverbs

Table 29 shows all attributes and values suggested by the EAGLES guidelines for adverbs. The attribute degree (i) is recommended, while the attributes (ii), (iii) and (iv) are optional.

Attribute	Values			
(i) Degree	1. Positive	2. Comparative	3. Superlative	
(ii) Adverb-Type	1. General	2. Degree	3. Particle	4. Pronominal
(iii) Polarity	1. Wh-type	2. Non-wh-type		
(iv) Wh-type	1. Interrogative	2. Relative	3. Exclamatory	

Table 29: EAGLES guidelines for adverbs

EAGLES attributes for adverbs not used in this tagset

Māori adverbs do not show grammatical degree (i), adverb-type (ii), polarity (iii) and wh-type (iv). Accordingly, these attributes are not taken into account in the intermediate tagset.

Intermediate tagset for adverbs

Table 30 shows the intermediate tagset for adverbs.

Description	Intermediate Tag	Example
Adverb	AV0000	aoake

Table 30: Intermediate tagset for Māori adverbs

4.9 Adpositions

The EAGLES guidelines propose one attribute for adposition, which is shown in Table 31.

Attribute	Values			
(i) Type	1. Preposition	2. Fused prep-art	3. Postposition	4. Circumposition

Table 31: EAGLES guidelines for adpositions

Type

For the attribute type (i), prepositions are the only type of adposition in Māori. Therefore the remaining values are ignored in the intermediate tagset.

Intermediate tagset for adpositions

Table 32 show the intermediate tagset for adpositions.

Description	Intermediate Tag	Example
Preposition	AP1	ko

Table 32: Intermediate tagset for Māori adpositions

4.10 Conjunctions

For conjunctions, the EAGLES guidelines propose three attributes as shown in Table 33. The attribute type (i) is recommended while the attributes (ii) and (iii) are optional.

Attribute	Values			
(i) Type	1. Coordinating	2. Subordinating		
(ii) Coord-Type	1. Simple	2. Correlative	3. Initial	4. Non-initial
(iii) Subord.-type	1. With-finite	2. With-infin.	3. Comparative	

Table 33: EAGLES guidelines for conjunctions

Type

With respect to the attributes type (i), Māori conjunctions can be categorized into two categories: coordinating and subordinating. However, due to a lack of an explicit and detailed description of Māori conjunctions, the attribute is ignored in the intermediate tagset.

EAGLES attributes for conjunctions not used in this tagset

The attributes coord-type (ii) and subord.-type (iii) are superfluous to Māori conjunctions. Accordingly these attributes are ignored in the intermediate tagset.

Intermediate tagset for conjunctions

Description	Intermediate Tag	Example
Conjunction	C000	erangi

Table 34: Intermediate tagset for Māori conjunctions

4.11 Numerals

For numerals, the EAGLES guidelines recommend the attributes (i)-(v) as shown in Table 35.

Attribute	Values			
(i) Type	1. Cardinal	2. Ordinal		
(ii) Gender	1. Masculine	2. Feminine	3. Neuter	
(iii) Number	1. Singular	2. Plural		
(iv) Case	1. Nominative	2. Genitive	3. Dative	4. Accusative
(v) Function	1. Pronoun	2. Determiner	3. Adjective	

Table 35: EAGLES guidelines for numerals

In some languages, numerals are not normally considered to be a separate part-of-speech because they can be subsumed under another category. Arguably, in Māori, cardinal numerals behave like verbs and ordinal numerals behave like adjectives (Harlow, 2001). Nonetheless, the EAGLES guidelines regard numerals as a major category. Note, the part-of-speech function of a word can be indicated by way of the attribute function (v).

Type

The attribute type (i) is relevant to Māori numerals. Like English, Māori makes a distinction between cardinal and ordinal numerals. Moreover, Māori ordinal numerals are morphologically marked by the prefix *tua-*.

Wh-type

The attribute wh-type (vi) is relevant to Māori numerals. Therefore a new attribute is added to the intermediate tagset, as shown in Table 36.

(vi) Wh-type	1. Interrogative	2. Relative	3. Exclamatory		

Table 36: An additional attribute wh-type for numerals

EAGLES attributes for numerals not used in this tagset

The attributes (ii)-(v) are ignored in the intermediate tagset because the attributes gender (ii), number (iii) and case (iv) are not relevant to Māori numerals.

Intermediate tagset for numerals

Description	Intermediate Tag	Example
Cardinal numeral	NU100000	rima
Ordinal numeral	NU200000	tuarima
Interrogative numeral	NU000001	hiag

Table 37: Intermediate tagset for Māori numerals

4.12 Interjections

The EAGLES guidelines do not propose any additional attributes for the class of interjections.

Intermediate tagset for interjections

Description	Intermediate Tag	Example
Interjection	I	kāti

Table 38: Intermediate tagset for Māori interjections

4.13 Unique/Unassigned

The EAGLES guidelines provide a unique category, intended for one-member word classes such as negative particles, existential particles and the infinitive marker (Leech, 1996). Although this category contains no recommend attributes, the EAGLES guidelines recognize that individual languages will need to identify such classes. Furthermore, the guidelines propose a single optional attribute unique-type (i) consisting of miscellaneous values, as shown in Table 39.

Attribute	**Values**			
(i) Unique-type	1. Infinitive marker	2. Negative particle	3. Existential marker	4. Second negative particle
	5. Anticipatory	6. Mediopassive voice marker	7. Preverbal particle	

Table 39: EAGLES guidelines for unique/unassigned

Unique-type

Concerning the attribute unique-type (i), special attention is to be paid to particles. Māori particles can be categorized into four categories: verbal particles, prepositions, determiners and postposed particles. Of these categories, verbal particles can be further subcategorized into preverbal particles and postverbal particles. Moreover, the EAGLES guidelines propose that prepositions and determiners are recognized as a separate part of speech, of which the intermediate tagset makes use. Of the three remaining classes, new values are defined for the attribute unique-type (i) as shown in Table 40.

Attribute	**Values**			
(i) Unique-type	1. Personal article	2. Negative particle	3. Preverbal particle	4. Postverbal particle
	5. Postposed particle			
(ii) Postposed	1. Manner	2. Directional	3. Locative	4. Other

part.-type				
(iii) Tense	1. Present	2. Imperfect	3. Future	4. Past
(iv) Aspect	1. Perfective	2. Imperfective		

Table 40: Additional attributes for Māori unique

Intermediate tagset for unique/unassigned

Description	**Intermediate Tag**	**Example**
Personal article	U1000	a

Table 41: Intermediate tagset for unique/unassigned 1

Description	**Intermediate Tag**	**Example**
Negative particle	U2000	kāhore

Table 42: Intermediate tagset for unique/unassigned 2

Description	**Intermediate Tag**	**Example**
Past preverbal particle	U3040	i
Perfective preverbal particle	U3001	kua
Preverbal particle	U3000	ka
Postverbal particle	U4000	ana

Table 43: Intermediate tagset for unique/unassigned 3

Description	**Intermediate Tag**	**Example**
Manner postposed particle	U5100	kau
Directional postposed particle	U5200	mai
Locative postposed particle	U5300	nei
Other postposed particle	U5400	anō

Table 44: Intermediate tagset for unique/unassigned 4

Description	**Intermediate Tag**	**Example**
Locative time	U6000	āpōpō

Table 45: Intermediate tagset for unique/unassigned 5

4.14 Residual

The EAGLES guidelines recommend three attributes for residuals. Residuals comprise of various semi-linguistic and non Māori elements as shown in Table 46.

Attribute	Values			
(i) Type	1. Foreign word	2. Formula	3. Symbol	4. Acronym
	5. Abbreviation	6. Unclassified		
(ii) Number	1. Singular	2. Plural		
(iii) Gender	1. Masculine	2. Feminine	3. Neuter	

Table 46: EAGLES guidelines for residual

Type

With respect to the attribute type (i), only the values foreign word and symbol are relevant to Māori. In the intermediate tagset, foreign words are textual elements that are not Māori, e.g. English words used in Māori texts. Furthermore, symbols are non-alphanumerical characters which are not punctuation.

EAGLES attributes for residual not used in this tagset

The attributes number (ii) and gender (iii) are not relevant to Māori. Accordingly, these attributes are ignored in the intermediate tagset.

Intermediate tagset for residual

Description	Intermediate Tag	Example
Foreign word	R100	English
Symbol	R300	*

Table 47: Intermediate tagset for Māori residuals

4.15 Punctuation marks

The EAGLES guidelines recommend two approaches for the mark-up of word-external punctuation. The first approach is to assign a separate tag for each main punctuation mark, e.g. period, comma, question mark, etc. The second approach is to group the punctuation marks into positional classes: sentence-final; sentence-medial; left-parenthetical; right-parenthetical. Needless to say, the second approach excludes potentially useful information. Therefore the first approach has been adopted for the intermediate tagset as shown in Table 48.

Intermediate tagset for punctuation

Description	Intermediate Tag	Example
Close parenthesis	PU1	)
Close quotation mark	PU2	”
Close square	PU3	]
Colon	PU4	:
Comma	PU5	,
Exclamation mark	PU6	!
Full stop	PU7	.
Neutral quotation mark	PU8	“
Open parenthesis	PU9	(
Open quotation mark	PUA	“
Open square	PUB	[
Question mark	PUC	?
Semi-colon	PUD	:

Table 48: Intermediate tagset for Māori punctuation

4.16 Māori Intermediate tagset

Table 49 shows the complete intermediate tagset for Māori.

Description	Intermediate Tag	Example
Interrogative common noun	N1000001	aha

Interrogative locative noun	N3000001	hea
Interrogative personal noun	N4000001	wai
Singular common noun	N1010000	wahine
Plural common noun	N1020000	wāhine
Common noun	N1000000	hātarei
Locative noun	N3000000	konei
Personal noun	N4000000	ākuhata
Noun	N0000000	
Transitive verb	V000000000000010	patu
Intransitive verb	V000000000000020	oma
Experience verb	V000000000000030	rongo
Neuter verb	V000000000000040	pakaru
Interrogative verb	V000000000000001	aha
Verb	V000000000000000	
Interrogative adjective	AJ00000001	pēhea
Adjective	AJ00000000	pai
First person inclusive dual personal pronoun	PD4030015000000	tāua
First person inclusive plural personal pronoun	PD4020015000000	tātou
First person exclusive singular personal pronoun	PD5010015000000	au
First person exclusive dual personal pronoun	PD5030015000000	māua
First person exclusive plural personal pronoun	PD5020015000000	mātou
Second person singular personal pronoun	PD2010015000000	koe
Second person dual personal pronoun	PD2030015000000	kōrua

Second person plural personal pronoun	PD2020015000000	koutou
Third person singular personal pronoun	PD3010015000000	ia
Third person dual personal pronoun	PD3030015000000	rāua
Third person plural personal pronoun	PD3020015000000	rātou
Demonstrative singular determiner	PD0010020100000	tēnā
Demonstrative plural determiner	PD0020020100000	ēnā
Interrogative singular determiner	PD0010020400000	tēhea
Interrogative plural determiner	PD0020020400000	ēhea
Singular determiner	PD0010020000000	taua
Plural determiner	PD0020020000000	aua
Possessive singular a-form determiner	PD0010020300001	tāku
Possessive plural a-form determiner	PD0020020300001	āku
Possessive singular o-form determiner	PD0010020300002	tōku
Possessive plural o-form determiner	PD0020020300002	ōku
Possessive singular neutral-form determiner	PD0010020300003	taku
Possessive plural neutral-form determiner	PD0020020300003	aku
Definite singular article	AT1010	te
Definite plural article	AT1020	ngā
Indefinite article	AT2000	hē

Adverb	AV0000	aoake
Preposition	AP1	ko
Conjunction	C000	erangi
Cardinal numeral	NU100000	rima
Ordinal numeral	NU200000	tuarima
Interrogative numeral	NU000001	hiag
Interjection	I	kāti
Personal article	U1000	a
Negative particle	U2000	kāhore
Past preverbal particle	U3040	i
Perfective preverbal particle	U3001	kua
Preverbal particle	U3000	ka
Postverbal particle	U4000	ana
Manner postposed particle	U5100	kau
Directional postposed particle	U5200	mai
Locative postposed particle	U5300	nei
Other postposed particle	U5400	anō
Locative time	U6000	āpōpō
Foreign word	R100	English
Symbol	R300	*
Close parenthesis	PU1	)
Close quotation mark	PU2	”
Close square	PU3	]
Colon	PU4	:
Comma	PU5	,
Exclamation mark	PU6	!
Full stop	PU7	.
Neutral quotation mark	PU8	“
Open parenthesis	PU9	(

Open quotation mark	PUA	“
Open square	PUB	[
Question mark	PUC	?
Semi-colon	PUD	:

Table 49: Intermediate tagset

References

Alqrainy, S. and Ayesh, A. (2005). “Developing a tagset for automated POS tagging in Arabic.” Centre for Computational Intelligence – School of Computing, De Montfort University, Leicester, U.K.

Baskaran, S. (2008). “Designing a Common POS-Tagset Framework for Indian Languages.” Microsoft Research, Bangalore, India.

Bauer, W. (1997). The Reed Reference Grammar of Māori. Wellington, New Zealand.

Biggs, B. (1998). Let’s learn Māori. Auckland, New Zealand: University Press.

Crandall, D. (2005) “Automatic accent restoration in Spanish text”. School of Informatics and Computing, Indiana University, Bloomington, U.S.

Hardie, A. (2003). “Developing a tagset for automated part-of-speech tagging in Urdu.” Department of Linguistics and Modern English Language, University of Lancaster, U.K.

Harlow, R. (2001). A Māori Reference Grammar. Auckland, New Zealand: Pearson Education.

Harlow, R. (2007). Māori: A Linguistic Introduction. Cambridge, United Kingdom: University Press.

Holmes, D. (1995). Māori Language: Understanding the Grammar. Dunedin, New Zealand.

Leech, G. and Wilson, A. (1996). Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG-TCWG-MAC/R.

Māori language Commission (2008). Hē Pātaka Kupu. New Zealand.

Mihalcea, R. (2002) “Diacritics Restoration: Learning from Letters versus Learning from Words”. Southern Methodist University, Computer Science and Engineering Department, Dallas, TX, U.S.

Scannell, K. (2010) “Statistical Unicodification of African Languages”. Department of Mathematics and Computer Science, Saint Louis University, St. Louis, Missouri, U.S.

Streiter, O. and Luca, E. (2003) “Example-based NLP for Minority Languages: Tasks, Resources and Tools.” European Academy, Bolzano, Bozen, Italy.

Tufis, D. and Ceausu, A. (2011) “Diacritics Restoration in Romanian Texts”. Institute for Artificial Intelligence, Romanian Academy.

Wagachar, P. and Pauw, G. (2006) “A Grapheme-Based Approach for Accent Restoration in Gikuyu”. School of Computing and Informatics, University of Nairobi, Nairobi, Kenya.

Williams, H.W. (2006). Dictionary of the Maori Language. Wellington, New Zealand.