Māori Vocabulary: A Study of Some High Frequency Homonyms

by
Kelly Elizabeth Keane-Tuala

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the requirements for the degree of
Master of Arts
in Māori Studies

Victoria University of Wellington
2013

Masters of Arts in Māori Studies

Māori Vocabulary: A study of some high frequency homonyms

Kelly Keane-Tuala

Abstract

The problem addressed in this thesis concerns the accuracy of Māori language vocabulary counts, e.g Boyce (2006), where Māori was found to use a very small vocabulary in comparison with e.g. English. As Boyce (2006, ii) acknowledges, this is partly explained by the degree of homonymy in Māori, which undermines the accuracy of the count. Homonymy is the phenomenon of the same string of letters (word-form) having two or more unrelated meanings (e.g. *kī* 'say', 'be full'). Automated word-form counts of Maori language texts count the form *kī* as the same word, regardless of its meaning. Unless different meanings of the same word-form are counted as different words, such counts will underestimate the vocabulary of the Māori language. (Homonymy is not the only explanation for the low count; further explanations have been suggested by Bauer (2009) and Nation (2011).)

The thesis explores whether there are consistent clues in the linguistic environment that signal the correct interpretation of homonyms in texts, and if so, how such clues could be used for tagging corpora so that counting would be more accurate. The Boyce corpus of modern broadcast Māori (Boyce, 2006, ii) provided the data. Case studies were made of three high-frequency homonyms in this corpus, *kī* 'say', 'full', *mea* 'say', 'thing' and *tau* 'settle', 'year'. Lyons' (1968) criterion of distinction was applied to establish the lexemes realised by each of these word-forms on the basis of dictionary and etymological information. The tokens of each word-form were then extracted from Boyce's (2006) corpus using the concordance program 'WordSmith Tools'. WordSmith Tools is a computer program that helps to look at how words behave in a text. Concord which is part of WordSmith Tools enables the user to see any word or phrase in context. Phrase peripheries (the words before and after each word-form in the same phrase) were analysed and the wider syntactic environment was also examined in order to find clues which signalled the appropriate lexeme for each token. The results showed that the lexemes from all three case studies could be identified in the corpus on the basis of consistent clues that occur in its linguistic environment. If the phrasal periphery of the word-form is examined, and the grammatical information supplied by the wider linguistic environment is taken into account, it is possible to determine the appropriate lexemic tag for a word-form in a corpus in Māori.

Acknowledgements

Ehara taku toa i te toa takitahi, he toa takitini

He mihi i te tuatahi ki te Atua, nāna nei ngā mea katoa.

He mihi i te tuarua ki a Winifred Bauer. He tohunga o te whatu toto o te reo Māori. Nāna au i akiaki kia whai tonu kia tutuki.

He mihi i te tuatoru ki ōku mātua mō ā rātou whāngai tamariki mutunga kore. Ki tōku whānau whānui mō ō rātou kaha ki te tautoko.

He mihi whakamutunga ki ōku toka tū moana, ki tōku hoa rangatira – ko Tai, ki āku tamariki – ko Joseph-Lee rātou ko Ezekiel, ko Leviticus mō ō rātou whakapono ki a au.

List of Abbreviations

I SG/PL	first person singular/plural
II SG/PL	second person singular/plural
III SG/PL	third person singular/plural
I DL INCL/EXCL	first person dual inclusive/exclusive
I PL INCL/EXCL	first person plural inclusive/exclusive
CLS	classifying particle
CONT	continuous particle
CONTR	contrastive particle
DET	determiner
DIST	distant
DO	direct object
EMPH	emphatic particle
EQ	equative
INTENS	intensifier
NEG	negator
N-FORM	nā/nō
PART	particle
PASS	passive
PERS	personal article
PL	plural
POSS	possessive
PREP	preposition
SG	singular
TAM	tense aspect mood marker
VOC	vocative

Contents

List of Tables

1 Introduction

> "English contains hundreds of thousands of words … In contrast, the Māori dictionary contains a mere handful of words" (Haden, 1992)

It has often been claimed that Māori is a vocabulary-poor language in comparison to English. These claims have been reinforced by Boyce (2006), which showed that in Māori, 200 different word types account for 82.4% of Boyce's corpus of modern broadcast Māori (Boyce, 2006, ii); in comparison, 2000 different word types account for about 80% of an English text (Nation, 2001:17). As pointed out by Boyce (2006:88) homonymy, the phenomenon of the same string of letters having two or more unrelated senses (e.g. *kī* 'say', 'be full'), provides some insight into the reason for such a low figure for Māori (although it is not the only explanation, see Bauer (2009) for other factors). This thesis is an examination of a few high frequency homonyms in Māori which aims to shed some light on the difficult question of how we might obtain more accurate counts of Māori vocabulary. To date, there are no tagging programs for Māori (John Cocks, personal communication). One of the reasons for this is the lack of overt morphology to provide clues for distinguishing the different meanings of a string of letters. The issue can be illustrated by the joke in example (1) which exemplifies the kinds of problems that arise due to the ambiguous nature of some lexemes in the language.

1.	He	aha	te	kī	a
	DET	what	theSG	key/say	of
	te	kūaha	e	kī	ana?
	theSG	door	TAM	lock/say/full	TAM
	"E	kī!	E	kī!"	
	VOC	say/key	VOC	say/key	

'What is the key to the door that is locked? Goodness me!'

'What does the door that speaks say? Hey key!'

'What does the door that is full say? Goodness me!'

The first person asks the question given in the first sentence of (1). The listener would probably assume that the question is 'What is the key to the door that is

locked?' and might answer 'The one that unlocks it?'.The joker then gives the answer to the joke - *E kī, e kī*! 'goodness me/hey key', and the listener realises that the question they have actually been asked is 'What does the door that speaks/is locked say?' or possibly 'What does the door that is full say?' which explains the response given by the joker "Hey key!" or "Goodness me!". Though there are many different meanings associated with each token of *kī* in example (1), all occurrences of *kī* would be counted as one and the same in a typical word counting program such as *WordSmith Tools*. Version 4 of this program was used by Boyce (2006) to count the tokens in the Māori Broadcast Corpus (MBC)

We will refer to each of the different meanings of *kī* as a separate lexeme so that *kī* 'say' is one lexeme, and *kī* 'key' is a different lexeme. The term lexeme will be explained more fully in Chapter 2. The first question this thesis attempts to answer is how do we get a more accurate count of the lexemes of Māori, as opposed to the word-forms? This thesis attempts to find information in the context of such word-forms which could potentially be used by a tagging program to disambiguate them, and thus enable them to be counted separately. In order to do that, it is first necessary to decide what the potential lexemes associated with a particular word-form are. This is discussed in detail in Chapters 2-3 where Lyons's (1968) criterion of distinction is used to establish lexemes on the basis of dictionary and etymological information. Having determined the potential set of lexemes associated with a particular word-form, the next part of the investigation involves determining which lexeme we have in any particular textual token. The resolution is different for grammatical particles (e.g *e* as in example (1)) and for content words. This thesis is concerned only with content words. The situation is different and slightly more complicated for grammatical particles and is discussed further in Chapter 3. In terms of content words, the words preceding and following the word are often sufficient to distinguish any particular textual token, but not always. Sometimes the wider grammatical context has to be considered. The implications of distinguishing lexemes based on the wider grammatical context are raised in the conclusion of this thesis.

Three Māori word-forms are examined in detail as the basis of this thesis:

kī ‘say’ ‘full’, *mea* ‘thing’ ‘say’ and *tau* ‘year’, ‘settle’, ‘lover/spouse’, ‘string of garment/loop’, ‘ridge of a hill’, ‘sing/song’, ‘attack’, ‘awesome’ ‘number’. The data from the MBC is used in gathering samples of the word-forms and also in the analysis of these word-forms. The case studies have been chosen for their high frequency, and because each poses somewhat different problems both for the determination of the appropriate lexemes they are associated with, and for the clues which allow recognition of the appropriate lexeme for any given token. Automated tagging programs are useful for tagging certain environments in Māori, which is discussed in areas for further research.

2 Terminology

2.0 Introduction

This section details the linguistic terminology used in the analysis of the data. There are two areas of terminology to explain: terms associated with morphological analysis, and terms associated with the grammatical analysis of Māori. The terms used in the grammatical description of the Māori language will be set out first in order to aid the discussion of the Māori examples used in this section. Then the terms used in morphology from a non-Polynesian perspective will be explained. Following this is detail of morphology and its implications for isolating types of languages.

2.1 Basic Units in Māori

“The phrase, not the word” (Biggs, 1969:17) in Māori is the most important unit for the discussion of Māori grammar. The following section describes those aspects of phrase structure in Māori which are important for the analysis of the data in this thesis. In this thesis, all examples which are not attributed to another source are my own.

2.1.1. The phrase

Phrases in Maori are of three kinds, firstly, a verb constituent (VC) as in example (1). A verb constituent according to Bauer (1997:12) is a phrase with a verb as its lexical head. Secondly, a noun phrase (NP), exemplified in (2), is a phrase that has a noun as its lexical head (Bauer, 1997:10). And lastly, a prepositional phrase (PP) in Māori always begins with a preposition and is followed by a noun phrase (Bauer, 1997:9).

1. kua mea
   TAM say
   ‘has said’
2. te mea
   theSG thing/say
   ‘the thing/the saying’

3. ki te mea
   to theSG thing
   'to the thing/ to the saying'

The phrase in Māori can be looked at in terms of the constituents that occur in it. The phrase is made up of phrase-type markers, lexical heads, and optionally one or more modifiers following the lexical head (but sometimes preceding it). These constituents of the phrase are described further below.

2.1.2. Phrase-type markers

Phrase-type markers indicate whether the phrase they introduce is a verb phrase, noun phrase or prepositional phrase. There are various phrase-type markers in Māori listed in Bauer (1997:8-9) and outlined below.

2.1.2.1 TAMs

Bauer (1997:8) states that Tense Aspect Mood Markers (TAMs) are particles which mark the phrase they introduce as a verb constituent, as with *kua* in example (4). A particle is a lexeme which has to be defined at least in part by its grammatical function. The glosses for these TAMs are suggestive only, and a fuller understanding of their meanings can be obtained from e.g. Bauer 1997, Chapters 6-9.

4. Kua tae mai te manuhiri
   TAM arrive hither theSG visitor
   'The visitor has (arrived)'

The TAM is important in the identification of those words that function as verbs in Māori and TAMs were key to differentiating lexemes in the data analysis. The TAMs listed by Bauer (1997:8) were regularly found in the data: *i* 'past', *kei te* 'non-past continuous', *i te* 'past continuous', *ka* 'relative past/present/future', *kua* 'perfect/inchoative', *e…ana* 'relative, present', *e* 'future/non-past', *ana* 'punctual/imperfective', *ai* 'habitual', *me* 'obligation', *kei*, 'monitory', *kia* 'subjunctive', and *ki te* 'infinitive'.

The TAMs that can be automatically tagged to identify the headwords that follow as verbal lexemes are *kia, ka* and *kua.* The remaining TAMs cannot be used to tag headwords that follow as verbal as their forms have other

functions and therefore are not reliable indicators. Some verb constituents in Māori do not have a TAM. In cases like this, as in example (5), the phrase is described as having a Ø TAM.

5. Mea atu au …
   say away ISG
   'I said …'

Verb constituents always function as the compulsory phrase in the predicate in a verbal sentence in Māori.

2.1.2.2 Determiners

Determiners mark the phrase that they introduce as a noun phrase. The following list of determiners is from Bauer (1997:8-9): *te* 'theSG', *ngā* 'thePL', *he* 'a', *a* (precedes a proper noun), *ia* 'each/every', *(t)ētahi* 'a/some', *(t)aua* 'that/those aforementioned', *(t)ēnei/nā/rā* 'this/these by me/by the listener/away from speaker and listener', *(t)ēwhea* 'which (sg/pl)'. The following are a few examples of a large paradigm of possessive determiners: *(t)ana/(t)āna* 'his/her(sg/pl)', *(t)ā rāua* 'their(dl; sg/pl)', *(t)ō kōrua* 'your(dl; sg/pl)', *(t)ā Rewi* 'Rewi's (sg/pl)' etc.

He 'a' is most commonly found as the head of predicative phrases in Māori. It is only classed as a determiner when it precedes nouns in subject noun phrases. The determiners *te* and *he* cause issues in the analysis of lexemes in Māori. This is explained further in section 2.6 under stem nominalisations.

Determiners are particularly useful when tagging parts of speech in Māori to identify nominal lexemes. The lexical head of the phrase in (6), *mea,* can automatically be analysed as a noun, making it distinct from the canonical transitive verb *mea* 'say' and the action intransitive verb *mea* 'say'.

6. ngā mea
   thePL thing
   'the things'

Noun phrases either occur as part of prepositional phrases, or function as subject noun phrases in Māori.

2.1.2.3 Prepositions

Prepositions introduce prepositional phrases in Māori and have noun phrases as their complements. Example (7) contains the preposition *ki* ‘to/at’, followed by the noun phrase *te kura* ‘the school’.

7. ki te kura
   to theSG school
   ‘to the school’

The prepositions that were most useful in distinguishing lexemes in this thesis were *i* and *ki*. When a verbal lexeme was accompanied by *i* marking a direct object or an adverbial expressing cause, and when *ki* introduced an adverbial of goal in my data, this provided an important clue to distinguishing between homonymous verbal lexemes.

Prepositional phrases function as adverbials or as the predicate in non-verbal sentences in Māori.

2.2 Lexical heads

The lexical head in Māori is the word that has inherent meaning, also referred to as a content word. The following lists the relevant classes of lexical heads in this study. If two content words occur, the first one is the lexical head, since modifiers follow the head in Māori.

2.2.1. Nouns

Bauer (1997:9) states that “the lexical head of a phrase with a determiner as phrase-type marker is a noun”. There are some cases, however, where a lexeme that is verbal in sense can occur as the lexical head of a phrase with a determiner as a phrase-type marker. These types of phrases are called stem nominalisations. Stem nominalisations are discussed in 2.6.

2.2.2. Verbs

The general definition of verb is the “lexical head of phrase with a TAM” (Bauer, 1997:9). There are several types of verbs in Māori. However, the only verb-types that will be discussed in this thesis are canonical transitive verbs, action intransitive verbs and state intransitive verbs. The other verb types in Māori,

such as experience verbs, neuter verbs and di-transitive verbs, are not found in my data and so will not be looked at further. The relevant verb types are characterised as follows:

2.2.2.1 Transitive verb

The type of transitive verb we are concerned with here is the canonical transitive verb. A canonical transitive verb most frequently co-occurs with a direct object phrase marked with the preposition *i* (Bauer, 1997:18) as in example (8). The subject noun phrase of these verb types is the actor or doer of the action and the direct object is the patient. The direct object of *mea* 'say' and *kī* 'say' is not always marked by *i;* this is discussed further in section 5.2.2.

8. Ka mea mai ia i tōna whakaaetanga
   TAM say hither IIISG DO his agreement
   'He will say that he agrees.' (more lit. 'He will say his agreement.')

2.2.2.2 Intransitive verb

There are two groups of intransitive verbs that will be examined in this thesis. Firstly, there are action intransitives in which the subject noun phrase expresses the actor or performer of the action, as in *ngā waka* 'the boats' in example (9). In contrast, there are state intransitive verbs where the subject noun phrase is found in a state identified by the verb. So for example, in (10) the state intransitive verb *kī* identifies the state 'full' that *ngā kete* 'the bags' are found in. In some respects, state intransitives in Māori are parallel to adjectives in English.

9. Ka tau ngā waka ki uta
   TAM anchor thePL boat to shore
   'The boats will anchor at the shore.'

10. Kua kī ngā kete i te kai
    TAM full thePL basket cause theSG food
    'The baskets are full of food.'

2.3 Modifiers

Modifiers are the final but non-obligatory part of the phrase in Māori. Bauer

(1997:16) states that “a modifier can be a single word or a phrase or a clause”. The types of modifier that we are mostly concerned with in the analysis of lexemes in this study are those that are single words and occur as verb modifiers in verb constituents.

2.3.1. Adverbial particles

Adverbial particles, for the most part, modify the head of a verbal constituent. Therefore, these particles are useful indicators signalling the verbal sense of a lexeme. In certain cases in the analysis of the data in this thesis, the following adverbial particles helped to distinguish lexemes.

2.3.1.1 Directional particles

From the point of view of my data some very important particles that function as verb modifiers are directional particles, especially when there is Ø marking of the TAM. The directional particles in Māori are: *mai* ‘hither’, *atu* ‘away’, *iho* ‘downward’, *ake* ‘upward’. The directional particles play an important role in differentiating verbal lexemes from nominal lexemes where they occurred in stem nominalisations (see 2.6 for stem nominalisations).

2.3.1.2 Manner Particles

The manner particles that were important in the identification of distinct lexemes in the data analysis were *tonu* ‘still’ and *kē* ‘instead’. These manner particles follow the head of verb constituents that they modify.

2.4 Obligatory Sentence Constituents

2.4.1 Functions of Phrases

Not only is the periphery of the phrase containing the lexeme examined in this thesis, but the role of the phrase in which the lexeme occurs is also examined.

There are two major types of sentences in Māori: verbal sentences and non-verbal sentences. The constituents for each type differ. Bauer (1997:5) explains that the constituents of sentences in Māori are phrases. All verbal sentences must contain a verb constituent, and no more needs to be said about that. There are three further phrase functions that we will look at here: predicate constituents, subject constituents and direct objects. It will be convenient to

consider the first two together.

2.4.2 Predicate constituents and subject constituents

Predicate constituents occur in non-verbal sentences. Example (11) has been broken into the predicate constituent [*he tamaiti pai* ‘a good child’] (referred to as the predicate phrase) and the subject constituent [*ia* ‘he/she’] (referred to herein as the subject noun phrase).

11. [He tamaiti pai] [ia]
    CLS child good IIISG
    ‘He/she is a good child’

The subject phrase is always a noun phrase in Māori, but predicate phrases may be either noun phrases or prepositional phrases. The latter thus have a number of different phrase-type markers. However, the most important phrase-type markers in relation to the analysis of my data are *he* and *ko*. Where a predicate phrase was marked with the phrase-type marker *he* or *ko* and where the homonymous lexeme occurred as the lexical head in this environment, the phrase was marked as predicate head in order to tag the lexeme for that particular environment. This was to distinguish between those lexemes that occurred in subject noun phrases and those that occurred as the predicate head of the predicate phrase. Bauer (1997:27) uses the term predicate phrase for the predicative constituent in non-verbal sentences in Māori and the term verb constituent for the compulsory predicative constituent in verbal sentences in Māori (Bauer, 1997:12). Where a lexeme occurred as the head of a predicate phrase or verb constituent in the data analysis, it was marked as predicate head.

2.4.3 Direct Objects

Direct Objects occur with transitive verbs but not with intransitive verbs in Māori. Direct objects are prepositional phrases marked with the preposition ‘*i*’. Due to the nature of the data analysed in this thesis, we are not concerned here with di-transitive verbs, nor are we concerned with experience verbs (the direct object of an experience verb is most often marked with ‘*ki*’). Since we are concerned only with canonical transitive verbs, all direct objects discussed here are marked with *‘i’*. Those direct objects marked with the preposition ‘*i*’ helped to

distinguish the lexeme *kī* ‘full’ from *kī* ‘say’ in Māori. Unfortunately it is not possible to automatically tag for these phrases in a corpus of Māori because *i* marks phrases of various kinds. This issue is discussed further in Chapter 8.

However, the canonical transitive verbs *mea* ‘say’ and *kī* ‘say’ in most cases co-occur with a direct quotation as their object and not an object phrase marked by *i*, as in examples (12) and (13).

12. Kua mea mai ia “He pai tērā”
TAM say hither IIISG CLS good that
‘She has said, “That is good”.’

13. I kī atu te māhita
TAM say away theSG teacher
“kaua e mahi kia pēnā!”
NEG TAM work TAM like that
‘The teacher said, “Don’t do it like that!”’

2.5 Adverbials

There are two types of adverbials that we are concerned with in this study. They are causer phrases and goal phrases.

2.5.1 Causer phrases

Causer phrases occur with state intransitive verbs and neuter verbs. We will only discuss those that co-occur with state intransitive verbs which are of relevance to this study. The causer phrase can be marked with either ‘*i*’ or ‘*ki*’. Bauer (1997:49) states that there is no evidence to suggest why one preposition is used over the other; however, *i* is the most commonly used phrase marker for causer phrases in my data. The high frequency of contexts where only *i* is possible is implied by Bauer (1997, 213-214) and also Harlow (2007, 156-157).

Example (14) shows the causer phrase *i te kai* ‘by the food’, which is marked with the preposition *i.* Causer phrases play a significant role in distinguishing the state intransitive lexeme *kī* ‘full’ from the canonical transitive verb *kī* ‘say’.

14.	Kua	kī	te	kete	i	te	kai
	TAM	full	theSG	basket	cause	theSG	food

‘The basket is full of food.’

2.5.2 Goal phrases

Goal phrases also play a significant part in the identification of distinct lexemes. The goal phrase is marked with the preposition *ki*. Bauer (1997:50) states that the goal phrase marks the end point of movement and only occurs following canonical transitives, di-transitives and action intransitives.

2.6 Stem nominalisations

As mentioned previously, stem nominalisations can cause issues in the analysis of the data in this thesis. On one level the principal lexeme in a stem nominalisation could be tagged as two distinct lexemes at the same time. So for example in most cases verb constituents can be tagged as an environment in which a verb will be found, the lexical head of a predicate phrase can be tagged as an environment in which a noun will be found, and the lexical head of a subject noun phrase can be tagged as a noun. However, stem nominalisations can express a verbal sense but occur in a predominantly nominal environment, i.e. in a predicate phrase or a subject noun phrase, or in a prepositional phrase as a direct object or adverbial.

Bauer (1997:524) states that stem nominalisations can be introduced by *te*, *he* or *hei* and that the degree in which they show nominal characteristics varies. Some lexemes in the lexical head of stem nominalisations can be considered more verbal than nominal despite the occurrence of *te*, *he* or *hei.* An example of a stem nominalisation introduced by *te* is as follows:

15.	te	tau	mai	o	ngā	waka	ki
	theSG	settle	hither	of	thePL	boat	to
	te	whanga					
	theSG	harbour					

‘the settling of the boats into the harbour’

We can see that in example (15) the gloss of *tau* is not nominal, but verbal in sense. The sentence is describing an action that is taking place, and so *tau* has

more of a verbal sense than a nominal one.

The issue that arises in the analysis of the data in this thesis is making the distinction between the lexeme *tau* ‘year’ for example, which is nominal and the lexeme *tau* ‘settle’ in a stem nominalisation, which is verbal. In most cases the word-form *tau* can be tagged as a noun when preceded by a determiner, and will thus be a token of the lexeme *tau ‘year’*. However if the determiner is *te*, *he* or *hei* it could be part of a stem nominalisation and therefore actually be a token of the lexeme *tau* ‘settle’. Where there was a lexeme that occurred as the lexical head of a stem nominalisation in the data in this thesis, it was marked as a stem nominalisation in order to look at whether the environment in which it occurred could have other clues to signal its verbal sense.

2.7 Morphology

We will begin by looking at morphology from a conventional linguistic perspective. According to Bauer (1988:4) morphology is concerned with firstly the identification of minimal meaningful units or morphemes. It is also concerned with their classification and the description of possible combinations in a convenient distributional unit, usually identified as the word (Bauer, 1988:7). The word, then, is generally regarded as the largest unit with which morphology is concerned.

2.8 The term ‘word’

According to Lyons (1968:403), “traditional grammar” was built on the foundational belief “that the word was the basic unit of syntax and semantics”. It is not an easy task to define exactly what the word ‘word’ means. This is the case both in lay usage and linguistically. In lay usage there are many subtle meanings of the word ‘word’ which will be discussed shortly, and linguistically there are many facets of a ‘word’. Therefore it is important to discuss these issues in order to come to an understanding about how it is defined within a metalanguage and how it is understood outside of this context. Once we have established the conventional terminology, we will then turn to how one might use the term ‘word’ in this study.

Firstly, we will look at what a word is and how it is defined linguistically. A

word, as defined by Lyons (1977:18):

> "is any sequence of letters which, in normal typographical practice, is bounded on either side by a space."

A similar definition is used by Bauer (1988:7) for the term 'orthographic word', that is, any word form bounded by spaces. Bauer uses other terms which are discussed and exemplified here. Firstly, we will look at how the lay person might identify a written word. Bauer (1988:7) discusses the sentence in example (16). He explains (1988:7) that in lay usage of the word *word*, one is most likely to reach the answer to the question 'how many words are there in example (16)?' by counting the items which occur between spaces on the page, therefore concluding that there are 15 words.

16. The cook was a good cook as cooks go, and as cooks go, she went
    (Bauer, 1988:7)

On one level, this answer is correct, and we can make it more precise by specifying that the sentence contains 15 orthographic words. (Bauer, 1988:7). However, some of these orthographic words are closely related, and the question arises as to how we identify and talk about these relationships. Bauer (1988:7) distinguishes between a word's features by using the terms 'word-forms', 'lexemes' and 'grammatical words'. The terms that are relevant to the morphological analysis of the Māori language are 'word-forms' and 'lexemes'. One question raised by Bauer (1988:7) is whether we say that *cook* is the same word as *cooks* and whether we see *go* and *went* as the same word? *Cook* and *cooks* are different 'orthographic words' (Bauer, 1988:7) and according to Bauer have different forms. Bauer (1988:7) introduces the terms 'lexeme' and 'word-form' to specify these relationships. All the occurrences of the orthographic word *cook* realize the same lexeme, which is written in small caps, COOK, and *go* and *went* realize the same lexeme GO. We can then say that *cook* and *cooks* are different 'word-forms' realizing the lexeme COOK, and *go* and *went* are different word-forms realizing the lexeme GO. There are thus four orthographic words in (16) which realize the lexeme COOK, and two different word-forms which realize COOK, namely *cook* and *cooks* (each of which occurs twice).

Given that the Māori language is more isolating, which of these distinctions do we need for Māori – and do we need others? Here we look at the

issues which arise when using the terms ‘word-form’ and ‘lexeme’ in Māori as they are understood in English. The first problem is the term ‘word-form’ and the difficulty of applying this term under Bauer’s description to Māori. In example (16) Bauer (1988:7) uses the singular word-form *cook* and the plural word-form *cooks* to show how different word-forms marking number realise the same lexeme. Consider example (17) where I have used the singular and plural of *kaitunu* ‘cook (noun)’.

17.

Pai	ake	te	kaitunu	ki	konei	i
good	INTENS	theSG	cook	at	here	than
ngā	kaitunu	ki	korā			
thePL	cook	at	there			

‘The cook here is much better than the cook over there.’

This example demonstrates the first significant difference between the Māori language and the English language. Where in English singular *cook* becomes plural *cooks,* which is created from the lexeme COOK by a morphological process in most cases, no parallel process applies in Māori. In order to produce the same change in number in Māori separate words are used. The phrase *te kaitunu* ‘the cook’ contains the singular ‘determiner’ *te* ‘the’, while the plural form *ngā kaitunu* ‘the cooks’ contains the plural determiner *ngā* ‘the’. In Māori, it is the determiners e.g. *te* and *ngā* that mark number in the noun phrase. Therefore, due to the difference in morphological process in English and Māori, the term ‘word-form’ is not as useful in Māori as it is in English. It should be mentioned here that there are a handful of nouns in Maori which do change form to mark number but are irregular forms, e.g. *wahine* ‘woman’ and *wāhine* ‘women’, *tangata* ‘person’ and *tāngata* ‘people’, *tamaiti* ‘child’ and *tamariki* ‘children’. The question is, is the term ‘word-form’ needed for such a highly isolating language as Māori? The answer to this question involves the inflection/derivation divide in Māori. We will return to this question soon.

Another exemplar of the differences between Māori and English is the marking of tense. In example (16) Bauer (1988:7) analyses *go* and *went* as being different word-forms which realise the same lexeme GO. Examples (18) and (19) provide a parallel example from Māori. If we consider the word *worked* in English, in accordance with Bauer’s criteria, we see that *work* and *worked* are

clearly different word forms of the lexeme WORK in English.

18. I mahi ia
    TAM.PT work IIISG
    'She worked'

19. Kei te mahi ia
    TAM.PRS work IIISG
    'She is working'

Examples (18) and (19) contain parallel examples of *mahi* 'work' in Māori. Employing the term 'word form' the question now is this: Is *i mahi* in (18) a 'word form' of the lexeme MAHI? We can see here that the present/past distinction does not lie in the word *mahi* 'work' at all but in the 'particle' *i* preceding it. The term 'particle' is explained further in the next section.

The notion of 'word-form' as understood in an English language context does not fit into the morphology of the Māori language as demonstrated here in singularity and plurality of words and marking of tense. So 'word-form' can be used for words like *te* and *ngā* and *tēnei* and *ēnei* but not for the same classes of words as in English.

Bauer (2003:14-15) demonstrates that the 'word form'/'lexeme' distinction is closely tied up with the distinction between inflection and derivation. It is to these phenomena we now turn and consider their application in a Māori language context.

English is usually described as having two types of morphology: inflection and derivation. The most significant principle of inflection as outlined by Bauer (2003:14) in relation to Māori is that inflection creates word-forms of a known lexeme. This principle is demonstrated in example (16) with the word-forms *cook* and *cooks,* where *cooks* contains the suffix *–s* which marks plural. This suffix is described as inflectional. Bauer (2003:14-15) again gives a broad explanation of derivation: derivation creates new lexemes from known lexemes, so for example the *–ess* in *goddess* and *priestess* is derivational because it creates new lexemes from GOD and PRIEST.

Bauer & Bauer (2012) discuss the inflection/derivation divide in Māori by taking several criteria into consideration. In this paper Bauer & Bauer looked at

all possible features of Māori morphology and using seven of 25 possible criteria drawn from Plank and other sources (2012:5), applied them to Māori in order to determine whether various morphological features in Māori were inflectional or derivational. The large number of criteria involved in making the divide between inflection and derivation gives an idea of the complexities involved. We will only look at the relevant parts to help us understand what the inflection/derivation divide reveals about the analysis of the Māori language. After applying their selected criteria to number marking on some nouns, marking on deictics, nominalisation marking, agentive marking, causative marking and passive marking, Bauer and Bauer found that the results appeared inconclusive. What would have been a clear-cut distinction in English was not so for Māori. Some of the criteria however are less useful for the analysis of Māori than they are for other languages. For example the criterion of change in word class does not apply as clearly in Māori as it does in English. It is normal for a word in Māori to occur as both the head of a noun phrase and as the head of a verb constituent, or many other types of environments. So, for example, in (20) *waiata* ‘sing’ functions as the head of the verb phrase and in (21) functions as the head of a predicate phrase. Example (22) exemplifies *waiata* ‘song’ as a modifier to *pukapuka* ‘book’. We will return to these types of environments soon.

20. Kei te waiata ngā tamariki
    TAM sing the$_{PL}$ children
    ‘The children are being good’

21. Ko te waiata tēnei
    PREP the$_{SG}$ song this
    ‘This is the song’

22. Kei hea te pukapuka waiata?
    PREP where the$_{SG}$ book song
    ‘Where is the song book?’

Bauer & Bauer (2012:5) state that the criteria used suggested that far more of the Māori processes examined had the properties expected of derivation processes than had inflectional properties, and they suggest that this could indicate that there is no clear distinction between derivation and inflection in Māori. On the other hand, they state (2012:5) that it could show that there is a

contrast there that is distinguished differently than it is in English. That the criteria do not align for Māori as they do in other languages could signal a need for different means of analysis for Māori.

As Bauer & Bauer (2012) state, not all of the criteria they selected were of equal value when applied to a Māori language context. Bauer & Bauer then conducted the analysis with only the stronger criteria for the Māori language, that is, the criteria of 'productivity' and 'agreement'. The analysis of these two criteria alone pointed at nominalisations and passives as being inflectional while all other processes seemed to be derivational. This method, however, excludes more than 20 of Plank's criteria for distinguishing between inflection and derivation, diluting the process considerably for the analysis of the Māori language.

What this then suggests is almost contradictory to the conventional understanding of inflection and derivation in English morphology. The results were not conclusive regarding the inflection and derivation divide in Māori but what the analysis did show was the complexities involved in using criteria not specifically devised for a Māori language context.

The effect that inflection and derivation has on the analysis of the data in this thesis is that if nominalisations and passives were indeed inflectional then all word-forms of the lexeme *kī* that are created by these processes would need to be extracted from the corpus and analysed as well. However, if these processes are derivational thus creating new lexemes from known lexemes, the analysis would not need to be altered. However, even if it were the case that nominalisations and passives were formed by the process of inflection and all other possible word-forms created by this process were included in the data analysis, the results would not change significantly as their frequency in the MBC is relatively low. For example, the word *tau* occurred 3,096 times in the MBC, those potential nominalised word forms that occurred in the MBC were *taunga* that appeared 49 times and *tauranga* which appeared 38 times.

The issue of an alternative analysis of Māori is not a new concept. Krupa (1982:43) and Biggs (1969:17) have acknowledged the issues associated with the analysis of the Māori language and offered other methodologies for dealing

with it.

Krupa (1982:43) discusses the issues which arise in all Polynesian languages when it comes to the description of the word and how it ought to be dealt with. In particular, he discusses whether the term ‘word’ ought to be used in the description of Polynesian languages at all. Krupa (1982:43) states:

> “…grammatical meanings are partly expressed within the framework of the word and partly within that of the phrase.”

Biggs cited in Krupa (1982:43) on the Māori language in particular states that:

> “The conventional division of linguistic descriptions into phonology, morphology and syntax runs into certain difficulty when the language being described is of an isolating type.”

An ‘isolating type’ of language according to Pawley (cited in Krupa (1982:44)) has words which consist usually only of a single morpheme. The discussion of the distinction between inflection and derivation sheds light on the reasons why the grammatical analysis of the Māori language becomes problematic when trying to fit it into the framework of non-isolating types of languages. The morpheme in Polynesian languages (specifically Māori) requires alternate analyses.

An alternate analysis into the syntax of the Māori language was developed by Bruce Biggs. Biggs (1969:17) states that:

> “The phrase, not the word, is the unit of Maori speech which must be emphasised in learning. It is the natural grammatical unit of the language, and even more importantly, it is the natural pause unit of speech.”

Biggs’ claim has influenced the types of terms which are now used in the grammatical analysis of the Māori language. Thus it must be concluded that due to the isolating nature of the Māori language, not all terms used in the description of other languages are appropriate for the description and analysis of the Māori language. Despite this there are terms that can be used in a Māori language context. The first term that will be used in this thesis to discuss words in Māori is ‘lexeme’. Lexemes are, according to Bauer (1988:8) the dictionary entries of words, not necessarily head entries, but you would expect to find their separate identities acknowledged in the dictionary. Therefore this study employs the system devised by Bauer (1997:2-21); the terms listed at the beginning of this chapter form the basis of the description of Māori in this thesis.

2.9 Homonymy and Polysemy

Lyons (1977:550) states that lexical ambiguity in languages is attributed to the phenomenon of either *homonymy* or *polysemy*. Lyons (1968:405) states that *homonymy* is the phenomenon of "two, or more, meanings" being "associated with the same form". The notion of *polysemy* is explained in Lyons (1977:550) as "one lexeme with several different senses". The terms 'meaning' and 'sense' are thus critical to the distinction between homonymy and polysemy, and therefore this section will begin by illustrating the application and relevance of these terms. Following this, the criteria for distinguishing between homonymy and polysemy will be explored following Lyons (1977:550). The implications of homonymy for the Māori lexicon will be be explained. A discussion about the aforementioned criteria from a Māori language perspective concludes this section.

2.10 Meaning

The object of this study is to explore ways to discriminate between orthographic words representing different lexemes, and this involves the discussion of word 'meaning'. Thus we need to examine the terminology for this area. There is detailed and rich terminology in the description of semantics in linguistics. We begin by investigating the term 'meaning' and draw on Aitchison (1987) and Lyons (1968) to explain.

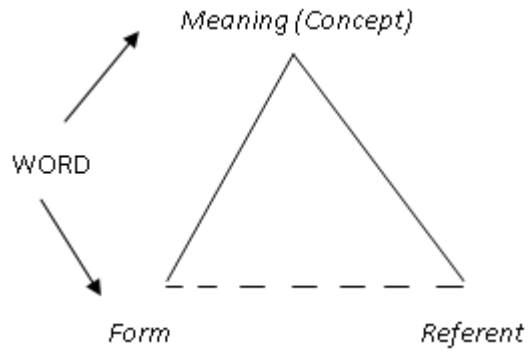
There are various theories about word-meaning. Lyons (1968:403) states that traditional grammar suggested that a word is composed of "two parts", firstly the 'form' ('form' understood as 'sign' or 'lexical item') and secondly its meaning. A distinction was made "between the 'meaning' of a word and the 'thing or things'" (Lyons, 1968:403) to which it referred. Throughout the history of traditional grammar the question arose as to what the relationship was between words and the things they referred to, or 'signified'. Lyons (1968:404) states:

> …the form of a word signified 'things' by virtue of the 'concept' associated with the form of the word in the minds of the speakers of the language; and the 'concept' looked at from this point of view, was the meaning of the word.

The term 'reference' was applied to the relationship between words and the

things that they “stand” for (Lyons, 1968:424). Within this notion of ‘reference’ Lyons (1968:425) explains that there are “pre-suppositions of ‘existence’”. This is inherent in an “ostensive” definition, that is, defining by “pointing to” the ‘referent’ or by indicating in some way (Lyons, 1968:424). Lyons presents these ideas diagrammatically in Figure (1).

Figure 1: Lyons (1968:404) ‘word-meaning':



Lyons (1968:404) defines *meaning* as the ‘concept’ of the object or *referent*.

The process by which the ‘meaning’ of a word is reached and then documented in the dictionary is as Aitchison (1987:43) discusses the process of determing ‘conditions of criteriality’ or ‘criteria attributes’, that is, the listing of necessary conditions in order to encapsulate the meaning of a word. Aitchison calls this the check-list theory which is used by most dictionaries when entering definitions of words. Regular exemplars of this theory are words like *square* and *bachelor* which have very distinct and fixed criteria attributed to them.

Atchison (1987:43) defines a square as satisfying these criteria: “it is a closed flat figure; it has four sides; all sides are equal in length; all interior angles are equal”. Philosophers like Aristotle argued that these ‘criteria attributes’ were appropriate and necessary in order to encapsulate the meaning of the word ‘square’. Similarly, *bachelor* has a fixed meaning, in that there is a limited set of criteria in order to establish what it is: HUMAN, MALE, ADULT, UNMARRIED. Unfortunately it is not always this easy to encapsulate a word’s meaning and there are certain words in languages that ‘criteria attributes’ cannot be applied to. One type of word in particular is ‘particles’ in Māori.

Particles in Māori are according to Bauer (1997:8), the ‘little words’ that are difficult to define. The particle *ko* in (23) cannot be defined in terms of its

meaning, but can only be defined in terms of its function in the sentence. *Ko* functions as a preposition and introduces some non-verbal sentences. The example in (23) exemplifies its use in an equational sentence. Bauer (1997:28) states that these types of sentences equate the subject i.e *te kaituhituhi* and the predicate phrase i.e *ko au*. Bauer (1997:28) also states that all equational sentences have predicate phrases introduced by *ko*. (There are other functions of *ko* which are not relevant to the discussion here.)

23. Ko au te kaituhituhi
    eq ISG theSG author
    'I am the author'

Particles of this kind cannot be defined in terms of their criteria attributes, as they do not have a *referent* in the same way that *'square and bachelor'* do. These types of particles are therefore defined in terms of their grammatical function.

In the analysis of lexemes of the Māori language in this thesis the word *meaning* will be used in reference to a *concept* signified by a *word-form* of a *lexeme.* We will look at some examples of how this applies to the Māori language shortly.

2.11 Sense

Belyayev (1963:145-147) explains the importance in language-learning contexts of teaching not only the meaning of a word but also its sense. Belyayev also states that the meaning of a word is "insufficient" in that there are usually multiple senses of words. These multiple senses can most accurately be remembered if "they are united in sense and are embraced in a general concept" (Belyayev, 1963:147). The idea of *sense* mentioned here by Belyayev is now examined more closely.

The idea of embracing a 'general concept' is best explained through an example. Consider the word *waka* in Māori which is often used as the Māori equivalent for 'car' in English. However, *waka* is traditionally the term for 'canoe'. Some, once aware that *waka* is the term for a 'canoe' may find it odd that this word is used for 'car'. These can two senses can be united by the common thread linked to Aitchison's (1987:43) 'criteria attributes', that is 'any

mode of man-made transport'. This general concept helps to unite all senses of *waka* 'canoe, vehicle, conveyance'.

Semantic fields can help to explain *sense* further. When placing words into *semantic fields* we intuitively understand that the words have a 'similarity of meaning' (Atkinson et al, 1982:179). Atkinson et al (1982:179) use the following sets of words to illustrate:

1. Cow, horse, tiger, animal, dormouse
2. Vehicle, car, bus, tandem, van
3. Chemistry, science, meteorology, physics, astronomy
4. Tree, forest, bower, wood, copse
5. Yellow, red, puce, violet, green

Grouping words into semantic fields gives an indication of a shared 'sense' between categories. In examples (1-5) we see that the general concept that each set of words shares is a 'natural' grouping such as (1) animals, (2) vehicles, (3) sciences, (4) 'woody' things (5) colours. These sets are *semantically similar* in that they all include a common concept. In these particular cases Atkinson et al were only interested in investigating *semantic fields* which were *semantically similar*. These are what they class as *paradigmatic relations*. There are issues which arise when using *semantic fields* regarding how broad or narrow one can be when making such lists. Words exhibiting *paradigmatic* relations are *semantically related* and all paradigmatically related words can occur in the same context. The words in the paradigm of *paradigmatic relations* have a related 'sense' yet differ in form *or* context of meaning or form *and* context of meaning (Lyons, 1968:428).

Atkinson et al (1982) also discuss the idea of words being "*semantically* related" rather than "*semantically similar*". An example of sets of words that are *semantically related* but not *semantically similar* are shown here in Atkinson et al's example (below) (1982:181):

Semantically related sets of words

a. Bark, dog

b. Mew, cat

c. Rancid, butter

The examples shown here in (a-c) demonstrate words which form a *syntagmatic* relationship. We understand the relationship between the first given word in each of (a-c) in relation to the second word by looking at all of them in context. Yet these words are not related on the same level as those words in a-e above, as they do not belong to the same syntactic class (Atkinson et al, 1982:181).

Another term which helps to explain the idea of *sense* is synonymy. Words which have a 'sameness of meaning' (Lyons, 1968:428) are regarded as being *synonymous*. If lexeme 'x' can be replaced with lexeme 'y' in a sentence and if the sentence maintains the same meaning once the substitution has taken place, then 'x' and 'y' are synonymous. Therefore *sense* and *meaning* in their non-technical usage themselves are synonymous. Lyons (1968:428) states that the "synonymy of lexical items is part of their sense". He goes on to say that "what we refer to as the sense of a lexical item is the whole set of *sense-relations* (including synonymy) which it contracts with other items in the vocabulary". Where *sense* is used in this study, it denotes the idea of the relationship a word-form or lexeme has with its meanings, and relationships between words with regard to both syntagmatic and paradigmatic relations. That is, a lexeme has different senses attached to it which relate to its meaning.

Let us now take a look at how the terms *meaning* and *sense* can be applied in a Māori language context.

The following are a selection of meanings and senses of *mea* drawn from (Williams, 1971):

a. Thing, fact, event, case, one

b. Say, intend, wish, think

c. Red, reddish

We will recognise three lexemes associated with *mea* with the associated meanings *thing, say* and *red*. Lexeme *mea 1* has the related senses 'thing, fact, event, case, one' and lexeme *mea 2* has the related senses 'say, intend, wish,

think', finally lexeme *mea 3* has the related senses 'red and reddish'. These will be discussed in terms of *meanings* and *senses.* The lexemes in examples (a-b) have distinct *meanings* attached to them but have several senses.

2.12 Homonymy and Polysemy Revisited

The phenomenon of a word form with multiple meanings is seemingly more noticeable and problematic because of the lack of morphology in Polynesian languages. We will begin by looking at the way the phenomenon of a word with multiple meanings is dealt with in standard introductions to the topic, and then we will look at what problems arise in a Māori language context.

The concepts of *homonymy* and *polysemy* are contested and controversial. The controversy surrounds how the distinction between these two concepts is made. Lyons (1968:405) discusses homonymy in relation to cases where "two or more, forms may be associated with the same meaning". These different meanings are usually the result of the two words having different origins. A word like *bank* for example means 'place where money is kept' and 'the side of a river, lake'. These are considered to be two different lexemes despite the fact that they have the same form because their meanings are different and unrelated.

Polysemy on the other hand, is generally used for words that have the same form, and which also have similarity of meaning, derivation or etymology. Aitchison (1994:60) describes polysemy as one lexeme with "multiple meanings" as does Lyons (1977:550-552). A polyseme is therefore generally described as a lexeme with 'multiple senses'. So how we do we decide whether we are dealing with homonymy or polysemy?

Lyons (1977:550-552) discusses three criteria that can be used to distinguish homonymy from polysemy. The methodology in this thesis follows a different chronological order to Lyons' ordering of the three criteria. The first criterion to be discussed is 'unrelatedness vs relatedness' of meaning (Lyons' second criterion). The second criterion to be discussed is etymology (Lyons' first criterion) and the third criterion is that of 'formal identity of grammatical function'.

The first criterion of distinction to be discussed here is ‘unrelatedness vs relatedness’ of meaning. This criterion is the only one proposed by Lyons (1977) that does not make use of diachronic information. Its drawback is that it relies on native speakers' intuitions with regard to the meanings of words and whether or not they are related. One common example used to illustrate the problems with this criterion is the word *‘ear’* as in the ‘body part’ sense and *‘ear’* as in the ‘ear of corn’ sense. Some native speakers naturally see (as Lyons, 1977:550 states) a metaphorical connection between the different senses of what they take to be the same word; they argue that there is a very obvious semantic relationship between one’s *ear* and an *ear* of corn, that is, that the shape of an ear of corn could be likened to the shape of the body part. Some native speakers insist that there is no relatedness between the two senses. So this particular criterion does not provide a resolution to the problem for 'ear', and the decision varies from one individual to another based on their own perception of the world around them. Lyons (1977:551-552) accepts this criterion of distinction and leaves the problems with it open to investigation.

The second criterion to be discussed from Lyons (1977:550) is ‘etymology’, that is, the history of a word and its origins. Etymology assists lexicographers in distinguishing between the phenomena of homonymy and polysemy. The words ‘*found’* and ‘*mouth’* will be used to illustrate. Lyons (1977:21-22) states that ‘found$_1$’ meaning “establish” and ‘found$_2$’ meaning “melt and pour into a mould” will be listed in most dictionaries as two separate entries with two distinct meanings. The words ‘found$_1$’ and ‘found$_2$’ exemplify two distinct lexemes by virtue of their etymology. “The historical derivation” of these two lexemes is that they come from “Latin ‘fundare’ vs. ‘fundere’”, which are “still distinct in modern French as *‘fonder’* vs. *‘fondre’*”. Thus these two English words (‘found$_1$’ and ‘found$_2$’) derive from historically different forms and are therefore to be analysed as different lexemes. By examining the origin of words and their original form it is possible to discriminate instances of homonymy from instances of polysemy. In contrast, the word ‘mouth’ is considered to be a polysemous lexeme with multiple senses attached to it, these senses being ‘organ of body’ and ‘entrance of cave’ and so on. Here we have one word-form that occurs with several different but related senses. The difference here in

contrast to '*found'* is that these meanings are of the same origin (Lyons, 1977:21-22, 550).

The third criterion to be discussed is the identity of grammatical function. "Homonyms" are generally defined as "lexemes all of whose forms have the same form" (Lyons, 1977:22). With reference again to the lexemes 'found$_1$*'* and *'*found$_2$*'*, both have the same set of forms *found, founds, founding* and *founded*. "There is identity of grammatical function", that is, "each lexeme is a verb" (Lyons, 1977:22) and is associated with the same set of inflectional forms. As previously discussed Māori has only a small number of word-forms when compared to a highly inflectional language like Latin. Grammatical function as a criterion from a Māori language perspective is not so much concerned with sets of forms, as with the grammatical function of each word in order to make the distinction between homonymy and polysemy.

2.13 Implications of homonymy for Māori

The importance of homonymy for this study of the Māori language is the way in which it affects word counts of the language. Automated word counts can only distinguish word-forms. This has different effects on word-counts of Māori and English.

English has significant numbers of inflected forms (with both derivational and inflectional affixes), each of which is counted separately. Māori, however, is an isolating type of language with few inflections, and an automated count thus underestimates the number of different words in Māori in comparison with English. To make a comparable count of Māori, lexemes must be counted separately.

This raises the question of how we know when we have distinct lexemes in Māori. Can lexemes in Māori be distinguished by their phrase peripheries and if so, how can this information be put to use to ensure more accurate counts? These are the issues that are addressed in the methodology developed for the case-studies that follow.

3 Methodology

3.0 Introduction

The chapter begins with a description of the Māori Broadcast Corpus (MBC) from which the data for analysis in this thesis has been extracted. The nature of the MBC is also discussed in order to shed light on the rationale for the chosen methodology. The criteria used for distinguishing lexemes in the corpus will be explained including a review of the dictionaries and also the etymological sources. Following this, the investigation of phrase peripheries and syntactic environments of lexemes is explained. A discussion of the contrast between lexical cases and grammatical cases is presented. Finally, the choice of case studies is explained in light of the methodology.

3.1 Description of the corpus

This section includes information about the MBC and how the methodology in this thesis is a response to the methods used and why.

3.1.1 The MBC

This research employs the work of Boyce and the invaluable compilation of Māori data in the MBC (Boyce, 2006). Boyce's work forms the foundation of this research which, without its existence, would not have been possible. This thesis attempts to refine the analysis of high frequency homonyms which are currently unaccounted for in her data. I also investigate whether it might be possible to tag effectively for these lexemes in corpora in Maori. This section looks at the work of Boyce and the methods I adopted in order to fulfil these aims.

The MBC – the basis of a PhD thesis by Mary Boyce – was published in 2006 and comprises modern Māori from radio broadcasts. It contains approximately one million words of running text (Boyce, 2006:6). The results showed that in Te Reo, 200 different word types account for 82.4% of Boyce's corpus of modern broadcast Māori (Boyce, 2006, ii); in comparison, 2000 different word types account for about 80% of an English text (Nation, 2001:17).

In the following passage Boyce (2006:100) discusses the results from the MBC which suggest that Māori is vocabulary poor.

> "The tables show that there are relatively few high frequency types in the MBC. This information on its own may be misleading. It may suggest that it would be a simple matter to learn these few word types and thus have full control of a large chunk of the language. It is by no means that simple. The information hidden behind the number of tokens of each word type is that a single word form, simply a series of characters bounded by spaces, may well represent a multiplicity of meanings." (Boyce, 2006:100-101)

Boyce exemplifies how the raw data can hide important information about the richness of words in her corpus by investigating the word *ana* (Boyce, 2006:108). She discusses the usefulness of the MBC as an aid for disambiguating its various senses. She points out that the corpus, can also be used to show the frequency of different uses of *ana* and other words in the MBC, and add new information not in current Māori dictionaries and grammars.

According to the information collated by Boyce (2006:108) there are four main uses of the word *ana*: it can function as a common noun, a post-posed particle, a possessive pronoun and a conjunction. Boyce (2006:109) showed that the function word uses of *ana* account for 99% of all tokens in her corpus. After randomly sampling 1 in 10 occurrences of *ana* as a function word, the MBC revealed that the postposed verbal particle use of *ana* when associated with pre-posed *e* was 95.09% of total function word uses; the possessive pronoun was 3.83% of total function word uses and 'other' which included the postposed verbal particle without *e*'.

The results above emphasise the importance of not only assessing whether the function word uses or content word uses are of higher frequency, but also the importance of disambiguating the senses of these various uses. The case with *ana* is one of many in which a word form encapsulates several senses.

Boyce (2006:108) mentions the problems which arise from the results in her corpus. Firstly, it gives a false indication as to what words students should be learning to be able to speak Te Reo. If students see Boyce's frequency list without the necessary context (http://tereomaori.tki.org.nz/Teacher-tools/Te-Whakaipurangi-Rauemi/High-frequency-word-lists) with no warning of

ambiguous forms then they are likely to learn fewer words and meanings than is necessary to competently speak and understand the language. Secondly the aim of her thesis was to provide a high frequency word list in Māori, and not to disambiguate word senses. Therefore, words with multiple senses are yet to be distinguished one from the other. These problems together increase the importance of my study which is aimed at finding out how to get a better count of lexemes with their distinct meanings and assessing which meanings are of higher frequency.

3.2 Choice of case studies

The first step in choosing the words for my case studies was to look at Boyce's list of the top 200 most frequent words in the MBC. I then explored the list for words that showed characteristics of homonymic lexemes, chosen according to their ranking in the MBC – the higher the frequency the more likely they were to be included in this thesis. Words that had separate head entries in dictionaries signalled possible homonymic lexemes. Those words that did not have separate entries but had seemingly unrelated senses – deduced by my own knowledge of Te Reo – were then selected for a short-list. The next step was to select from that list words that would contrast in the issues they were likely to raise for the investigation. So for example, in the dictionary review *kī* was assigned the grammatical label state intransitive verb and canonical transitive verb. The case study of *kī* looks at differentiating a verb from a verb. *Mea* was labelled as a canonical transitive verb, an action instransitive verb and a noun, and therefore we can investigate the properties of homonyms in different syntactic categories here. *Tau* however had homonymous lexemes including more than one noun and more than one verb. The case studies then provide a very broad investigation with regard to differentiating homonymous lexemes that are both similar in form and grammatical function or similar in form but different in grammatical function.

3.3 Annotating the data

Atkins (2008:254) discusses the FrameNet project in which an online lexical resource was built. The corpus data examples were manually annotated in

order to encapsulate the range of “syntactic and semantic combinatory possibilities (the valence) of a word in each of its senses” (Atkins, 2008:254). Atkins (2008:254) states

> The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions…

Therefore the above steps provided suitable lexemes for this thesis. All the tokens of each of these lexemes were then extracted from the MBC using *WordSmith Tools*. A total of 10,879 examples of *kī, mea and tau* were extracted from the data and the examples were then annotated for various environments, both syntactic and semantic, then sorted into categories in an Excel spreadsheet. Appendix (1) shows a random sample of the annotated data for each case study. Each case study was then assigned codes in columns. The codes were: Stem nominalisation (stem nom); type/fixed forms; Sense; Prepositions (prep); TAM; Determiner (det); Pre-posed Modifier; Modifier (mod); Sentence/Clause position; Collocates; Notes.

Where I could see from the context that the example was indeed a stem nominalisation Yes (Y) was inserted in the stem nom column. Types/fixed forms were to tag for examples of a specific use of a word in the text, so for example where *mea* ‘thing’ was used as a substitute for a person, place or thing, this was then annotated with ‘substitute’; this then gave an idea of how many of these uses were occurring in the MBC. Tagging for this particular use is not necessarily valuable toward distinguishing it in a text, but here it served the purpose of a partial explanation toward the high frequency of *mea* in the MBC. The next code was Sense: every one of the 10,879 examples was tagged for its specific sense as used in the MBC. The following codes: Preposition, TAM, Determiner, pre-posed modifier and post-posed modifier were to tag for the phrase peripheries. Where a preposition occurred preceding a lexeme, the preposition was then written in the column marked ‘prep’. This was the same process for TAMs, Determiners and Modifiers. Sample analysis pages can be found in Appendix 1.

3.3.1. The spoken nature of the corpus

The spoken nature of the MBC poses several problems for the analysis of the data undertaken in this thesis. The first issue can be seen by the contrast with a written corpus using published material which would be edited for things like hesitations and false starts. Even in scripted spoken material, these types of phenomena occur. There were many examples in the MBC where hesitations could be found. One example of quite an impressive chunk of hesitation from the MBC is given by Bauer (2009).

1. Pea, *kua* mutu *pea*. E, *e* huri *ana aku* mahara, *te* aroha ki ngā, *ki ō tātou* teina, tuakana o, o, *o* Poihākena, ā ngā nē, i, *i* wini *rātou* te, *te* tikanga mō te, ā, mō. Ā, *engari*, ā, te, *taku* pōhēhē, kua, *kua* mutu te, ā, kua, *kua tū* te, ā, *te mea*, *te mea nui*, ā, te, te, *nē*. Kua, *kua* mate *pea*, ā, te, te *e* huri *ana ngā* mahara *ki a rātou* ngā, *ngā* tēina, *ngā* tuākana *o* Poihākena.

Here I have replaced Bauer's colour coding with different typographical formats. The underlined function words are hesitations and the italicised ones are those that Bauer (2009) considered to be part of the actual message. The point here is that Māori speakers use function words like *ā*, *o* and *e* etc as hesitations, whereas English uses words like *um* and *er* etc. A program like WordSmith can be told to tag for words like *um* and *er* and remove them. Because Māori speakers use function words for hesitations, a program like WordSmith could not tag for these occurrences and remove them without also removing non-hesitations. The second point to be made about such hesitations is that the unedited data in the corpus affects the accuracy of word counts of high frequency words. The corpus is riddled with hesitations and it was rare to find a whole paragraph without hesitation. All the underlined words in example (1) were counted as uses and, although redundant, were tallied against that word-form which in turn affected the overall frequency of every word-form and this leads to inaccurate frequency numbers.

Another issue with hesitation is that whole phrases can be used in a hesitation (as in examples (4)-(6) below where repeated phrases are in square brackets). Examples (2) and (6) are particularly ambiguous, because we have an example of the phrase *te mea* used emphatically in sentence (2) and we have an example used as a hesitation in example (6). It can thus be difficult to

determine whether such instances as in (2) – (6) are cases of repetition for hesitation or repetition for emphasis. Without access to the original spoken data, deciding whether phrasal repetition was an emphatic use or just a hesitation was time-consuming. Where there was doubt, I consulted a native speaker; if this speaker was doubtful I then discarded the example from the analysis. Examples (4-6) were deleted from the analysis. The repeated phrases have been translated in order to give context to the phrase involved. Other parts of these examples are not glossed both because they are irrelevant to my point, and because it is often unclear what is intended.

2. rohia i runga i te ririrangi [te mea], [te mea], [te mea].
   '[etc], [etc], [etc]'
3. te tohungia [ko mea noa iho], [ko mea noa iho].
   '[whoever it may be], [whoever it may be]'
4. hī tūhāhā, tūhīhī. [Ka tau], [ka tau]. [Kua hoki]. [Kua hoki] ki te
   [settle], [settle]. [Have returned] [have returned]
5. kua maha ngā tau, , [kua maha], [kua maha] ngā
   [Have been many], [have been many]
6. ā, ko tēnei [te mea], [te mea], [te mea], te taetanga ki te Pākehā
   [the thing], [the thing], [the thing]

The spoken nature of the MBC also influences the use of certain lexemes. For example, the lexeme *mea* 'thing' is used in hesitations like those in examples (7) and (8). Example (7) illustrates the use of *mea* in place of a thing and (8) is in place of a person. The high frequency of *mea* 'thing' could possibly have been influenced by the spoken nature of the corpus as there are almost 500 uses of *mea* that were used either as a hesitation or in place of a person, place or thing while the speaker searched for the correct word. Of course superficially similar examples can be found in written material such as example (3) when the person, place or thing is irrelevant to the discussion, but these written cases are not signs of disfluency, unlike those in the MBC.

7. engari kāore he, [he mea] he ārai mō ngā, mō ngā rauemi nei …
   'but there wasn't any, um, [things], screens for the, for these resources.'

8. Ā, e, e te whaea, e mea, ā, te whānau nei, , ā …
   'Um, by, by the mother, [by who], um, this family, um, …'

3.3.2. Competence and performance

The difference between competence and performance is another issue we face with this type of corpus. The competence/performance distinction is postulated here by Noam Chomsky (1965:3-5):

> We thus make a fundamental distinction between *competence* (the speaker – hearer's knowedge of his language) and *performance* (the actual use of language in concrete situations). (Chomsky, 1965:3)

The competence/performance distinction is about the gap between what a speaker knows and what a speaker actually does. Where there is no script involved performance errors are bound to occur. There were a lot of examples of strange sentence structures in the MBC that were clearly ungrammatical. One instance here is example (9) where there is no verb following the TAM *kua.* There are a couple of possibilities to explain what could be responsible for the omission. It could possibly be a typo and the transcriber actually left the verb out, or if it was a native speaker talking, they might have filled the verb slot with a shrug, and assumed the listener could supply the appropriate verb.

9.	Kāre	mā	ngā	pākehā	nei,	i	te
	NEG	belong	thePL	pakeha	here	by	theSG
	mea	kua	rātou,	nē.			
	thing	TAM	IIIPL	Q			

'It wasn't according to these pakeha, because they have, eh?'

Examples of this kind sometimes had to be excluded from the analysis, if the ungrammaticality affected the word-form under consideration.

3.3.3. Omissions due to spoken discourse

In my own experience with spoken Māori it is common in informal speech for particles to be omitted and then left for the listener to understand from context. As a result, the expected items in phrase peripheries, such as phrase-type markers, do not occur, and so are not available to facilitate the tagging process for verb constituents, noun phrases and prepositional phrases. There are

dialects such as the Ngāti Porou dialect that that omit object markers (Bauer 1997:150) and omitted phrase-type markers also occur in written texts, so this type of occurrence is not solely restricted to a spoken corpus. However a spoken corpus would certainly include a higher occurrence of this phenomenon. There were hundreds of examples of omitted TAMs, determiners and prepositions in all three case studies. However, as explained in the results section of each chapter, it was deduced from the findings that if there was no directional particle or TAM *ana* following the lexical head, the lexeme was most likely nominal. There were cases where there was no directional particle or TAM *ana* as in (10) and (11). Example (10) is ambiguous because the modifiers *pai* 'good', *aroha* 'loving' could co-occur with both a nominal lexeme and a verbal lexeme. The adverbial expressing goal *i runga i te rangimārie* is most commonly used with action intransitives. The meaning could be either 'well settled, lovingly settled, settled on peace' or it could be 'a good year, a year of love, a year established on peace'. Example (11) contains *mea* 'say' with Ø TAM; examples of this kind pose problems for tagging this type of corpus for *mea* 'say', because the most useful clues, namely a TAM or a post-posed verbal modifier, are both absent.

10.	Tau	pai,	tau	aroha,	tau	i	runga
	settle	good	settle	affection	settle	PREP	on
	i	te	rangimārie				
	PREP	theSG	peace				

Settle well, settle in affection, settle under the mantle of peace.

11.	Ka	tau	mai	ngā	matawaka
	TAM	settle	hither	thePL	kinship group
	[mea	ōi]			
	say	shout			

'the kinship groups settled here and [shouted]'

3.3.4. Transcriber error

There are clear cases where some words have not been transcribed correctly. For instance example (12) contains the word *tau*. At first sight, it looked as though it might be the 'settle' sense. Yet the words in its environment did not

appear to be appropriate collocates of ‘settle’. After further investigation with the help of an informant, it seemed that the appropriate transcription should have been *tou* ‘bottom’. An example of its correct use is taken from *Te Kohinga o Wharekura* (Learning Media Limited, 2010) in example (13) where we find *he tou whiore kē te pākiwaha nei* ‘the big mouth was a coward’. This is certainly a case of transcriber error.

12.	kei te	āhua		tau	whiore	rātou
	TAM.PRT	somewhat		bottom	tail	IIIPL
	ki	te	whakatakoto		kōrero	
	at	theSG	lay down		say	

13.	he	tou	whiore	kē	te	pākiwaha
	DET	bottom	tail	instead	theSG	braggart
	nei					
	PART					
	[The big mouth was a coward.]					

This error can probably be traced to change in the vowel system of Māori, the issue that the MAONZE project has been investigating over the last 10 or so years. Speakers of Māori born as early as the 1930’s have been influenced by the articulation of the vowel system from English. The MAONZE project has shown that these two diphthongs (/ou/ and /au/) have become almost identical in modern Māori, with /au/ the normal pronunciation. Thus the form traditionally written as <tou> is typically pronounced /tau/ today, which probably explains the transcriber error here (Harlow et al; 2009:142).

Another example of transcriber error is the placement of, or omission of punctuation. Omission of punctuation is illustrated in example (14). There should have been a comma placed after *tau* in this example. What has been transcribed here is a *karakia* ‘incantation/prayer’ in Māori. A different text of the same *karakia* from (Salmond, 1975:161) reproduced in (15) shows that this is a poetic use of *hā*. The lack of comma wrongly suggests that *tau* is being used as some kind of pre-posed verbal modifier or that *hā* is a modifier to *tau*.

14.Tihe uriuri, tihe nakonako. Ka **tau** hā, whakatau, whakatau

15. Ka tau, hā, whakatau ko te
TAM settle ha place PREP theSG
papa i raro nei, ka tau,
earth PREP below here TAM settle
hā, ko Te Mataku mai i Rarotonga,
ha PREP Te Mataku hither PREP Rarotonga
'[It] lay, ha, set its place the earth below, trace back from Mataku from Rarotonga'

3.3.5. Elimination of Unusable data

Where there was clearly non-fluent speech that I could not sensibly classify, the data was discarded. Also where an example could not be assigned a sense from the context, it was discarded. The examples of *mea* used as a substitution for a person, place or thing were kept in the data (this accounted for almost 500 examples). The reason for keeping these examples in the data was because it was a word that had a function in the context of the spoken data. This use was also recognised by most dictionaries which validated its occurrence. There were 373 examples that were discarded from the analysis of *tau* leaving a total of 2723 sentences. A total of 322 examples were discarded from the analysis of *mea* with a total of 5740 remaining. There were 68 examples of *kī* that were discarded leaving a total of 1653.

3.4 Determining Lexemes

Part of the process of annotating the corpus data was to determine the appropriate lexemes for each case study. The following is a discussion of how I went about determining lexemes.

3.4.1. Lyons' criteria

As outlined in Chapter 2, Lyons (1977:552) discusses the criterion of distinction as a means of differentiating between homonymy and polysemy. Lyons' (1977:552) idea of 'relatedness vs unrelatedness of meaning' relies on native speakers' intuitions as to whether multiple senses of the same word form are related or unrelated. The lexicon in the dictionaries is representative of native speakers' intuitions and has been used as a guide as to how these multiple

senses are recognised. This part of the investigation gives a preliminary idea of how many lexemes there might be for a particular word-form. The purpose of the following dictionary and etymological review is to determine a method of deciding how many lexemes of *kī*, *mea* and *tau* there are. The purpose of establishing those lexemes is to consider whether lexemes can be distinguished in a corpus on the basis of their grammatical features.

3.4.2. Dictionary Review

There were several criteria that were important for the selection of dictionaries used in this review. Firstly, the dictionaries needed to be representative of native speakers' intuitions in accordance with Lyons' criterion. Secondly it was important to analyse both traditional Māori and modern Māori dictionaries. Finally, the authors of the dictionaries must be reputable.

The following gives the background of each dictionary and the ways in which they aligned with the selection criteria for a well-balanced representation of the Māori lexicon.

He Pātaka Kupu was created specifically for speakers of Māori. This resource is a monolingual Māori dictionary. It gives a good indication as to what Māori speakers consider to be separate lexemes and has head entries for lexemes it considers to be distinct. It became apparent, however, while using this dictionary as a reference, that it replicated a lot of information from Williams (1971). Though it was fully written in Māori, all entries and sub-headings were exactly the same as Williams. Therefore, *He Pātaka Kupu* was only included in the case studies where it had contrasting sense(s) to Williams. *He Pātaka Kupu* was not consistent with its grammatical labels, for example, all four dictionaries used in the review agreed on three labels for *mea* 'thing/ say', namely indefinite pronoun (IPN), transitive verb (VT) and noun (N). However *He Pātaka Kupu* lists action intransitive verb (VI) and state intransitive verb (VS) as added labels. There was no clear indication as to which grammarian they were basing their grammatical classifications on. Another example was that *mea* 'red/reddish' was only given as a lexeme by Williams and *He Pātaka Kupu*. Williams listed the grammatical function as VS, whereas *He Pātaka Kupu* listed it as VI and N. Therefore, where *He Pātaka Kupu* did not list different senses to Williams I did

not include comment on any grammatical labels because their labels were inconsistent and did not fit any type of model that I was familiar with.

Williams (1971) was an essential guide for this thesis as it contains traditional sources of information as well as offering a wider range of lexemes with examples. First printed in 1844 with the seventh edition in 1971, it is the foundation lexicon for the Māori language. Biggs (1990) was also representative of a traditional lexicon and both Biggs and Williams have been referred to in scholarly work such as Bauer (1997), Boyce (2006) and Harlow (2006) to name a few. Moorfield's (2003) *Te Aka* represents a modern lexicon in Māori. Moorfield is a specialist in Māori language, literature and culture. Included in his publications are a series of four graduated textbooks and resources teaching Māori to teenagers and adults called *Te Whanake* which is widely used as a resource by tertiary institutions. Moorfield (2003) is available online and is continuously growing.

Tirohia/Kimihia (2006) was designed for the learner and the teacher of Māori in kura kaupapa Māori (Māori immersion primary schools), and therefore its target audience is children aged between 8-12 years. Compiled in accordance with the results from the MBC it was published by Huia Publishing and is a monolingual Māori dictionary. This dictionary is also representative of a modern lexicon.

The one consistent theme with regard to the dictionaries in this review is that there is a difference in opinion as to what constitutes a lexeme and therefore what should be listed as a head entry. The boundary between homonymy and polysemy in Māori dictionaries is not as clear-cut as one might hope. The presentation of words in dictionaries can often be misleading. Kilgarriff (2008:143) states that lexicographers are often presenting grey areas of interpretation of senses that is senses that are the same and senses that are different are not given a clear division. He goes on to discuss that each lexicographer works within his own framework which is influenced by multiple things and states,

> The division of a word's meaning into senses is forced onto lexicographers by the economic and cultural setting within which they work. Lexicographers are obliged to describe words as if all words had a discrete, non-overlapping set of senses. It does not follow that they do, nor that lexicographers believe that they do. (Kilgarriff (2008:143)

Another issue that was prevalent in the dictionaries was that some lexicographers kept to a minimalist's lexicon listing only those words that they considered to be high frequency words and not listing words that are of a low frequency. Biggs (1990), Moorfield (2003) and *Tirohia Kimihia* (2006) are all dictionaries that had a grammarian either compiling or helping to compile the dictionary and these are the three minimalist lexicons as shown by the results of the dictionary review in each case study.

Lyons (1977:551-552) also discusses the issues surrounding English examples such as 'port' and 'ear' where native speakers' intuitions may be misleading in providing an accurate account of such words. Therefore further justification of the division into lexemes can be provided by looking into the etymology of a word.

3.4.3. Etymology review

Etymology is the next criterion used by Lyons in order to distinguish homonymy from polysemy. The discussion of etymology will be developed with reference to Tregear's work in the *Māori – Polynesian Comparative Dictionary* (Tregear, 1891) and the work of Clark and Greenhill in *The Polynesian Lexicon Online* (POLLEX). These are the only works of their kind which provide an insight into the history of the Māori language. These sources form the foundation for investigating lexemes and their relationships in the Polynesian language group. The lexemes of *kī, mea*, and *tau* listed by Tregear are looked at and subsequently compared with cognates found in Polynesia which he considers to be related lexemes. This analysis is then compared and contrasted to those reconstructed forms cited in Greenhill & Clark (2011). This work will help to establish whether the multiple senses of *kī, mea*, and *tau* are to be analysed as distinct lexemes or as a single lexeme with multiple senses by virtue of the etymological background. The information sourced from POLLEX includes reconstructed forms from within a sub-group only if they share the same form

and sense. This contrasts with Tregear who lists forms from higher nodes of the family tree even if the form or sense differs. Tregear states that his work provides the reader with those Polynesian words which are related to the Māori dialect. Tregear's work provides information about how the lexemes from Māori can be traced up the language family tree to Proto-Polynesian.

3.5 Structures

Once the previous steps have determined the appropriate lexemes for each word-form, and their associated meanings, the grammatical features of these lexemes are examined in order to establish the type(s) of environment(s) in which each lexeme occurs. This process helps to specify a set of syntactic guidelines for the environment(s) of each lexeme. This determines the basis for grouping the data into syntactic categories, consequently enabling a tagger to tag for each distinct lexeme in a corpus of Te Reo.

The results section for each case study provides information from the data within the MBC and looks at the frequency of each sense and the environments in which it occurs. Although most cases could be tagged for their syntactic environment, there are other very important factors that could not be tagged for computationally in a corpus. These include things like animacy or inanimacy of the subject; collocates which belong with distinct lexemes; direct objects, and adverbials expressing cause and goal.

The results include two-tailed P value results indicating which factors were significantly different for the lexemes under consideration. The two-tailed P values were calculated using the on-line tool found at http://graphpad.com/quickcalcs/chisquared1/. For each phrase-periphery item, this process compares the proportions of that item occurring with each of the focus lexemes with the proportions of the word-form which are attributed to each of those focus lexemes. The two-tailed P value requires the use of a null hypothesis (H0), and the results of the two-tailed P value analysis are used to determine whether the null hypothesis is retained or discarded. The Null Hypothesis developed for this study is that "There will be no significant difference between the percentages of a word form attributed to each associated lexeme, and the percentages of a word in the phrase periphery co-

occurring with each associated lexeme. Thus to conform to the null hypothesis, if a word-form W has two associated lexemes L1 and L2, where 80% of the occurrences of W are L1 and 20% are L2, then 80% of the occurrences of a particle p in the phrase periphery of W should occur with W=L1, and 20% should occur with W=L2". The Null Hypothesis was discarded where $\chi 2$ for each collocate was greater than the critical value determined by the degrees of freedom (number of lexemes minus one) and a confidence level of 95% probability of the Null Hypothesis being disproved. The results are discussed in the results section of the case studies in the thesis. The online tool provided an assessment of the level of significance of each P value, ranging from 'extremely significant' to 'not significant', and these interpretations have been included with the statistical results.

The final analysis and results section of each case-study chapter details how a tagger might tag for each sense of the word-form concerned. This step provides a reliable diagnostic for corpus analysis of the data in this thesis.

3.6 Lexical vs grammatical word-forms

3.6.1. Function words vs. Content words

The three case studies presented in this thesis are based on content words. Content words are those words that are more readily defined in terms of their meaning. Function words however are difficult to define in terms of their meaning and are instead associated with their function. The function words in Māori are particles, which were discussed in Section 2.10. A case like *e* for example has eight possible functions and therefore eight very different environments in which it occurs. Although function words are of higher frequency in the MBC, the disambiguation of these forms becomes problematic in the division between their grammatical labels, as exemplified in the discussion of *e* that follows.

3.6.2. The particle e

I considered the particle *e* as a possible word for a case study in this thesis. It is ranked as the 6th most frequent word in the MBC and has a number of different functions. These functions include: TAM future, present and non-past and in the

presence of the post-verbal particle *ana* also ‘continuous’. *E* also has a vocative function preceding personal nouns and pronouns with two morae or fewer, and occurs before intransitive imperatives under the same phonological condition. *E* also precedes numerals 2-9, and is the preposition which precedes agent noun phrases in the passive construction. After looking for the different functions of *e* I could fairly readily distinguish the preposition use in passives and the TAM use using information about the items that follow the particle. However, the issue of any sub-classes of the TAM *e* raised far-reaching problems, for example whether the *e* before 2-mora imperatives and the *e* before numerals 2-9 are TAMs or not. (See Bauer (1997:450-458) Therefore investigating this type of environment was going to lead me too far astray from the topic at hand and thus I decided to confine myself only to content words for the purposes of this thesis. That being said, I do believe that in most instances, the grammatical context – although not the phrase periphery – would provide clues for grammatical particles as well.

4 Kī

4.0. Introduction

The aim of this chapter is to determine what lexemes are realised by the word-form *kī,* and then to look at contextual clues surrounding the lexemes (which turn out to be *kī* 'full' and *kī* 'say') in order to tag these words in a corpus as distinct lexemes. To achieve this, my steps are firstly to investigate the dictionary meanings for each lexeme. The purpose of this is to look at how various dictionaries identify these words and whether or not they are recognised as distinct lexemes. This will provide insight into whether the speaker of Te Reo identifies these words as being either homonymous or polysemous. This chapter will also look at the etymology of *kī* 'full' and 'say' as documented in Tregear (1891) and Greenhill & Clark (2011), its grammatical features as presented in grammars and the results from the analysis of the raw data from the MBC. The contextual clues that differentiate the lexemes *kī* 'say' and *kī* 'full' in a corpus will be explained.

A tabulation of the information from the selected dictionaries precedes the discussion.

Table 4.1 Information from dictionaries about *kī*

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Tirohia Kimihia***	***He Pātaka Kupu***
Kī¹ Say	√	√	√	√	√
Associated grammatical functions:	*VT (-ia)*	*VI, VT, N*	*VI, VT*	*VI, VT*	*VI, VT (-ia,), N (-nga -anga)*
Senses:	*'tell, ask'*	*VI 'speak'VT 'say, call, mention, tell, designate'N 'saying/word'*	*VI 'speak'VT 'tell of, mention, call, designate, consider, think, imagine, speak, utter a word'*	*VI 'speak'VT 'say'*	*VI 'speak'VT 'ask, tell, explain, call/name, describe, imagine'N 'saying/word'*
Kī² *Full*	√	√	√	√	√
Associated grammatical functions:	*VS*	*VS*	*VS*	*VS*	*VS, N (-nga)*

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Tirohia Kimihia***	***He Pātaka Kupu***
Kī³ Key		√		√	√
Associated grammatical functions:		*N, VT, VS*		*N*	*N*
Senses:		*N 'key', 'lock'VT 'to lock'VS 'to be locked'*		*'key'*	*'key'*
Kī⁴Very			√		
Associated grammatical functions:			*MOD*		
Senses:			*'very'*		

4.1. Establishing lexemes associated with *kī*

4.1.1 Dictionary analysis of kī

The following is a selection of information from the reputable dictionaries mentioned in the previous chapter. The information has been placed into sections in Table 4.1. Each column is headed with the dictionary name, and the information from each dictionary has been placed in the following order: meaning; associated grammatical functions and finally, senses. The associated grammatical functions are abbreviated as follows: Noun (N), Transitive Verb (VT), State Intransitive Verb (VS), Action Intransitive Verb (VI), Modifier (MOD).

The Grammatical Function row indicates its function according to that particular dictionary. The grammatical function row also indicates the passive suffix (if any) which attaches to that particular form (according to the dictionary). The passive suffix has been included in order to distinguish sense, that is, it is useful to help differentiate the sense of the word forms where no grammatical label has been given. For example, refer to *He Pātaka Kupu* sense number (3) of the lexeme *kī*[1] in Table 1.1. This sense of the lexeme has not been assigned a grammatical label in the dictionary. We can infer an appropriate grammatical function by looking at the affixes associated with that sense. The sense has to reflect the passive suffix *-ia* and that the nominal suffixes *-nga* and *-anga* are associated with it. However information is not supplied by all dictionaries; some list the nominalisations only when they are semantically unpredictable, i.e. lexicalised.

Where a dictionary has distinguished separate lexemes, they have been assigned separate rows. For example, if the word-form *kī* has three separate meanings associated with it in a particular dictionary they are represented by *kī*$_1$, *kī*$_2$, *kī*$_3$ etc. For comparability the lexemes in the table are labelled as *kī*$_1$ (‘say’), *kī*$_2$ (‘full’), *kī*$_3$ (‘key’) and *kī*$_4$ (‘very’) where each is a distinct lexeme. Since the dictionaries vary in the numbering of the senses, this has been standardised in Table 4.1. If there is a tick (√) in the rows labelled ‘*kī*$_1$-*kī*$_4$’ this indicates that the dictionary in the column heading recognises this sense in its entry for *kī*. The column labelled ‘sense’ shows the various senses which have been assigned by the dictionary to the lexeme. Where a cell has been left blank there is no

information available of this kind.

From Table 4.1 it can be seen that *Moorfield*, *Tirohia Kimihia* and *He Pātaka Kupu* identify three lexemes associated with the word-form *kī*. The meanings attached to those three lexemes are ‘say’, ‘full’ and ‘key’. Williams also distinguishes between three separate lexemes, however the meanings associated with these lexemes are ‘say’, ‘full’ and ‘very’. Biggs (1991) lists only two lexemes, that is, ‘full’ and ‘say’.

The grammatical functions given by these various sources are as follows: the meaning ‘key’ has been classed as a noun by all dictionaries listing it, and also as a state intransitive when used in the ‘to be locked’ sense. All dictionaries agree with the function state intransitive for the sense ‘full’. However, *He* Pātaka *Kupu* (HPK) is the only dictionary to add the affix *–nga* under the entry for ‘full’. HPK also labels ‘full’ as a noun. Williams (1971) has classed the sense ‘very’ as an adverb.

According to the dictionaries reviewed here we have four likely lexemes associated with *kī* in Māori, that is, ‘say’, ‘full’, ‘key’ and ‘very’. The lexeme *kī* ‘full’ has no other senses associated with it by the dictionaries concerned. The information that we can glean from this review of *kī* are those lexemes that native speakers consider to be associated with this word-form.

To confirm that these are the appropriate lexemes, it is also necessary to consider their etymology. A tabulation of the etymological information precedes the discussion of the etymology.

Table 4.2 Senses of *kī* listed by Tregear (1891:145) and Greenhill & Clark (2011)

Tregear		*Greenhill & Clark (2011).*	
1.	*a. full:* *b. makikī, filled up; tight* *c. wharekī, a parent of many children (a "full house")* *d. high (of the tide)*	*1.*	*be full* *a. of container, house,* *b. stomach etc.*
2.	*very*		
3.	*not; not yet;*		
4.	*to say;* *a. to think; to speak, to utter a word;* *b. speech, an address: whaiki, to make a formal speech;* *c. pākiki, to question urgently;* *d. whaki, to confess.*	*2.*	*a. say* *b. speak* *c. tell* *d. mention* *d. call* *e. think*

Table 4.3 Polynesian lexemes and meanings associated with *kī* 'full' listed by Tregear (1891:145) and Greenhill & Clark (2011)

	Tregear		***Greenhill & Clark (2011)***	
PN Language	***Lexeme***	***Sense***	***Lexeme***	***Sense***
Samoan	*ʻiʻo*	*full, as a bottle or well; full-sized.*		
Tahitian	*i (ì)*	*full:*	*'ii.*	*plein, rempli*
Hawaiian	*ii*	*a gathering together*	*ki*	*full*
Marquesan	*cf. kikina*	*full.*	*ki*	*oko ki, très fort, superlatif*
Paumotan	*ki*	*full, replete; fakaki, to heap up; to fill; replete.*	*kii*	*to be full*
Rarotongan	*ki*	*to fill*	*kii*	*be full, fill (intr.), teem; fullness, bounty, wealth*
Penrhyn			*kii*	*covered, filled up, full*
Pukapuka			*kii*	*full*
Rennellese			*kikii*	*magino kikii, dead calm (of the sea)*
Vaeakau-Taumako			*ki/kii*	*depth*

Table 4.4 Proto-Polynesian lexemes and meanings associated with *kī* 'say' Tregear and Greenhill & Clark (2011)

PN Language	***Tregear***		***Greenhill & Clark (2011)***	
Samoan	*ʻi*	*to cry, as a fly or a bird; ʻiʻi (ʻiʻi), to give a prolonged scream or squeak.*		
Tahitian	*i*	*to speak (obsolete)*	*i*	*to speak (obs.)*
Tongan	*i*	*a. to squeak; kiki, to squeak;* *b. chickens; faka-kiki, to scream, to squeak; to make a shrill noise;* *c. to affright.*		
Hawaiian	*ki*	a. *to speak, to say* *b. to address one, to make a formal speech;* *c. to say within oneself, to think;* *d. to pronounce a single word as a signal;* *d. to give an appellation; ii, a rejoicing with an audible voice, like a chant*	*'Ii*	*to say, speak; suppose; saying*
Mangarevan,	*ki*	*to believe; to imagine; to think.*	*ki mai, ki atu*	*dire*
Easter Island			*kii*	*to say, speak*
Manihiki-Rakahanga			*ki*	*to say*
Moriori			*ki*	*say*
Rarotongan			*kii*	*talk, speak (Bse)*
Tuamotu			*kii*	*to speak, say; voice, word (n)*

4.1.2 Etymology

We can look further at the word *kī* in other Polynesian languages and see what the data might suggest with regard to how many lexemes *kī* there are on the basis of etymology.

Tregear (1891:145) has four entries for the word *kī* as listed in Table 4.2 whereas Greenhill & Clark (2011) has only two. The senses of *kī* in Māori in Table 4.2 are, according to Tregear (1891:145) and Greenhill & Clark (2011), separate lexemes which are related to the lexemes in Polynesian languages in Table 4.3 and Table 4.4. The senses given in Table 4.2 from Tregear (1891:145) added a further lexeme *kī* to the list derived from the dictionary review, since no dictionary listed 'not; not yet'. Greenhill & Clark (2011) does not regard the sense 'not' as a separate lexeme, which is consistent with the senses from the dictionary review. Moorfield (2011) and Williams (1971) note the use of *kīhai anō* and *kīanō* where the form *kī* is combined with *hai* or *anō* to form a single unit. The sense 'not' can be found only in combination with *anō* and/or *hai*. Where the sense is dependent on a set combination of words in a phrase it is not considered a single lexeme but is considered a multi-word unit. Therefore this sense of *kī* as 'not' will not be included in the analysis as a single lexeme by virtue of its meaning being reliant upon the unit.

Table 4.3 is a comparative account of the Polynesian lexemes considered to be related to the lexeme *kī* 'full' in Māori. These are sourced from Tregear (1891:145) and Greenhill & Clark (2011). Where a language derives from French Polynesia, POLLEX supplies the original French gloss. Greenhill & Clark (2011) gives examples which reconstruct to Tahitic languages and Tregear covers a wider range and more langages across Polynesia. The table shows agreement between POLLEX and Tregear with regard to the languages Tahitian, Hawaiian, Marquesan and Rarotongan as having lexemes related to *kī* 'full' in Māori. Though Biggs gives lexemes which construct to Tahitic languages, also included are Marquesan, Rennellese and Vaeakau-Taumako languages which sit outside of the Tahitic language family. The languages which have been flagged as problematic by Biggs are Marquesas, Pukapuka, Rennellese and Vaeakau-Taumako. The most likely reasons for marking examples as problematic are phonological irregularity, and dubious semantic

connection. Where the majority of languages maintain the 'full' sense of *kī* those problematic languages retain less similar sense and in some cases have a different form, such as Rennellesse *kikī* 'dead calm'.

4.1.3 Lexeme summary

The information from Māori dictionaries and the Māori – Polynesian comparative dictionaries has provided some justification for the following to be considered distinct lexemes: *kī* 'full', *kī* 'say' and *kī* 'key'. Williams (1971) allotted *kī* 'very' its own entry, considering it a separate lexeme. Tregear (1891:145) also considered *kī* 'very' to be a separate lexeme. However, *kī* 'very' did not appear in the corpus data. This suggests this lexeme is of low frequency and so is not included in my study. As mentioned earlier, the word *kī* 'not, not yet' was also omitted from consideration due to the fact that its sense is most regularly found in Māori in a multi-word unit – *kīhai* and *kīanō*. It can also be mentioned here that there are no instances of *kī* 'not, not yet' in the MBC. This tells us that it is rarely used in spoken Māori nowadays.

The words *kī* 'full' and *kī* 'say' can be justifiably analysed as distinct lexemes by virtue of their etymology. In Tables 4.3 and 4.4 we see that there are clear uses of *kī* 'full' and *kī* 'say' across Polynesia and if we track the data back through the Polynesian family tree, we can see the languages within the proto-Tahitic family have the same form for both senses of the word. When we move up the tree to the Proto-central-eastern node of the family tree the form is again the same. If we go another node up the tree to Proto-nuclear-Polynesian, we see that the form for the sense 'full' is different to that of the 'say' sense. SAM *'i'o* 'full' and SAM *'i* 'say' are different forms. We now have evidence to support the claim that *kī* 'say' and *kī* 'full' are distinct lexemes in Māori, because they have different ancestors. This evidence is also supported by Lyons' second criterion of distinction 'relatedness vs unrelatedness of meaning' which relies on native speakers' intuitions which were represented by the dictionary review.

4.2. Grammatical review

In order to facilitate the discussion of the syntactic analysis of the word *kī* it is useful to look at the grammatical functions and how each word has syntactic constraints that can signal the correct sense of *kī*. Drawing on Biggs' (1969:17)

claim that it is the phrase and not the word that is the smallest meaningful unit in Māori, we see that the word *kī* appears in a variety of phrase types, but the phrase types differ from one lexeme to another.

4.2.1 The grammatical functions of kī *'full'*

Kī 'full' has various grammatical labels assigned by various grammarians such as stative verbs or stative adjectives. However, here we use the label state intransitive employed by Bauer (1997:14). State intransitive verbs are a sub-group of intransitive verbs. An intransitive verb requires only one participant, that is, a subject constituent. A state intransitive can be likened to an adjective in English, yet typically in Māori these words also function as verbs. The nature of the state intransitive is that the subject constituent of the verb phrase in which it occurs is a patient rather than agent or actor. In contrast, the subject constituent of an action intransitive verb is the actor or agent.

1.	E	kī	ana	te	wai	nei	i
	TAM	full	TAM	the	water	here	by
	ngā	tuna					
	thePL	eel					
	'The water is full of eel'						

Example (1) shows some features of the syntax of a state intransitive sentence. The subject constituent *te wai nei* is stated to be in a state of being full. State intransitive sentences may optionally include a causer phrase which is marked with the preposition *i*. This is the function of the prepositional phrase *i ngā tuna.*

If present, the causer phrase can be important in distinguishing the two meanings of *kī*. The subject noun phrase also plays an important role in signalling the right interpretation of *kī*. This will be further developed in Table 4.10

4.2.2 The grammatical functions of kī *'say'*

There are two environments in which *kī* 'say' occurs. *Kī* 'say' may be a transitive verb where the subject constituent is the agent and the direct object is the thing being said. Thus in example (2), which is a more typical canonical transitive verb, the agent phrase is *te whāea* and the patient or direct object of the canonical transitive verb *whāngai* is *tāna pēpi* which is found in the prepositional

phrase marked with *i*. Compare example (2) with example (3) where the direct object of *kī* ‘say’ is *kei te whakaae ahau* which is a clause as opposed to a prepositional phrase marked by either *i* or *ki*. The canonical transitive verb *kī* ‘say’ can have a variety of structures as the direct object. Example (4) shows the direct object in the prepositional phrase marked by *i.* However it is still the thing being said. Bauer (1997:18) states that DOs do not function in the same way that other phrases with the same form do. Therefore it is important to distinguish between canonical DOs and other *i*-phrases. Because the direct object of *kī* ‘say’ in example (4) has a similar form to the cause phrase in (1), this is an example of the possible ambiguity between a cause phrase and a direct object. However the directional particle *mai* signals the correct interpretation here. There are cases of the directional particle *atu* occurring as a post verbal modifier to *kī* ‘full’, therefore instances like this would need to be further differentiated by collocates that co-occur with the lexeme ‘full’.

2. Kei te whāngai te whāea i tāna
   TAM feed theSG mother DO herSG
   pēpi
   baby
   ‘The mother is feeding her baby.’

3. Kei te kī mai te tangata
   TAM say hither theSG person
   “Kei te whakaae ahau”
   “TAM agree ISG”
   The person is saying “I agree.”

4. Kei te kī mai te tangata i tōna
   TAM say hither theSG person DO hisSG
   whakaaetanga
   agreement
   ‘The person is stating his agreement.’

The lexeme *kī* ‘say’ also has the sense ‘speak’. When the sense ‘speak’ occurs, *kī* functions as an action intransitive verb. An action intransitive verb logically has one participant, that is, the agent, but no direct object. Example (5) shows the sense of *kī* ‘speak’ with no direct object. The prepositional phrase *ki ngā*

manuhiri is an adverbial phrase expressing the goal.

5.	Kei te	kī	atu	te	kaikōrero	ki	ngā
	TAM	say	away	theSG	speaker	to	thePL
	manuhiri						
	guest						

‘The speaker is addressing the visitors.’

The dictionary review demonstrates some inconsistency between grammatical labels and a word’s function. For instance, Moorfield and *He Pātaka Kupu* note *kī* ‘say’ as a noun with the sense ‘saying/word’, yet the remaining dictionaries do not. Grammarians disagree on the range of functions assigned to the lexemes. It is again noted that Bauer’s term ‘stem nominalisation’ for these types of environments is employed here. Stem nominalisations are discussed in 2.6.

4.3. Results of *kī* from analysis of MBC

Kī occurs 1,747 times in the MBC and accounts for 0.17% of all tokens. It is the fifth most frequent homonymous item in the MBC. After the analysis of *kī* was complete, three senses of *kī* surfaced from the raw data in the MBC. Those three senses were: ‘say’, ‘full’ and ‘key’. The sense ‘say’ had the highest frequency, followed by ‘full’ and then ‘key’. The following table shows the frequency of these items.

Table 4.5 Raw frequency results for senses of *kī*

	'say'	'full'	'key'	**Total**
No. of Tokens	**1,653**	**41**	**1**	**1695**

The following section outlines the process of analysis of the established lexemes *kī* 'full', *kī* 'say' and *kī* 'key'. It will also look at the different environments in which *kī* occurs. An attempt is made to establish a set of syntactic guidelines to identify these three distinct lexemes in a corpus, thus enabling a tagger to effectively tag for each distinct lexeme.

Firstly the grammatical features of the verbal lexemes *kī* 'full' and *kī* 'say' will be examined in order to establish those environments where these words systematically appear. Secondly, the functional categories will be analysed, such as the phrase-type markers that each lexeme co-occurs with and their relative distribution; the modifiers that occur in the phrase and their co-occurrence with each lexeme; and the sentence position of each lexeme will also be looked at in order to conclude what the position might tell us about the sense of the word. Those items which are not easily distinguished by the previous two steps will be further analysed by the wider context in which they occur and also by word collocates which occur in the subject noun phrases and prepositional phrases where relevant. The results of these processes will then be discussed and final conclusions about how successful these processes were in distinguishing each lexeme will be made.

4.3.1 Grammatical features associated with each lexeme in MBC

In the grammar review of the lexemes *kī* 'say' and *kī* 'full' the types of environments in which these words occur grammatically were observed. The grammatical features of each lexeme act as the first clues in distinguishing their sense. The dictionary review established the following syntactic features for each lexeme: a state intransitive for the lexeme *kī* 'full' and both action intransitive and canonical transitive for the lexeme *kī* 'say'.

Every dictionary in Table 4.1 apart from Biggs distinguishes both action intransitive and canonical transitive uses of *kī* ‘say’. In the grammar review it was noted that the grammatical functions are useful indicators in the identification of the sense of a lexeme.

The sense ‘full’ adheres to the properties of a state intransitive verb as outlined in Bauer (1997:14). Williams (1971) and Biggs (1990) list *kī* ‘full’ as an adjective, but although the grammatical label is different, the function is similar to that of a state intransitive. As mentioned in the grammar review, there are similarities between state intransitives in Māori and adjectives in English. So we now consider the state intransitive sentence and what its environment can tell us about its sense.

A state intransitive sentence may also contain an adverbial with the preposition *i* which expresses cause. In example (6) the subject *ngā puku* is in a ‘state’ of ‘fullness’ caused by the means expressed in the adverbial *i te kai*.

6.	Kua	kī	ngā	puku	o	ēnei	tama	i
	TAM	full	thePL	stomach	of	these	boy	by
	te	kai						
	theSG	food						

‘The children’s stomachs are full of food.’

As mentioned in the grammar review an adverbial expressing cause in a state intransitive sentence where *kī* ‘full’ occurs and the DO of the transitive lexeme *kī* ‘say’ are key when distinguishing between the two lexemes and form the basis of tagging for the two uses. Example (7) contains the negative sentence *kāore au e haere* which functions as the DO of the canonical transitive verb *kī* ‘say’. Example (8) shows another canonical transitive sentence in which the subject noun phrase was omitted under identity and which contains the active sentence *kei te hē* which functions as the DO of *kī* ‘say’.

7.	Kī	atu	au	“Kāore	au	e	haere”
	say	away	1SG	not	1SG	TAM	go

‘I said, “I’m not going.”’

8.	E	kī	ana	“Kei te	hē”
	TAM	say	TAM	TAM	wrong

‘[They] said, “[It] was wrong.”

The data taken from the MBC did not consistently mark the DO of *kī* ‘say’ in inverted commas. This is one of the issues of the nature of the corpus which would hinder the process of tagging for this sense. If the DO of *kī* ‘say’ was consistently supplied with inverted commas in a corpus, a tagger could look for phrases marked with inverted commas in the environment of *kī* thus signalling the ‘say’ sense. Cases where the DO was not marked with inverted commas did cause issues identifying its sense, whereas this could have been rather an easy task.

If we compare the canonical transitive sentence structure to that of a state intransitive we see that the preposition *i* plays an important role in the distinction between the two. Examples (9) and (10) were easily identifiable as the sense of *kī* ‘full’ due to the adverbial which followed the verb constituent introduced by *i* as in *i te kai* in (9) and *i te kapu tī me te kai* in (10). Since these do not express utterances or summaries of utterances (e.g. ‘his agreement’, ‘these words’), they are very unlikely to be DOs of the canonical transitive *kī* ‘say’.

9.	Ka	kī	te	kāpata	i	te	kai
	TAM	full	theSG	cupboard	cause	theSG	food

‘The cupboard will be full of food.’

10.	Kī	atu	i	te	kapū	tī	me
	full	away	cause	theSG	cup	tea	with
	te	kai					
	theSG	food					

‘[It] was full of the cup of tea and the food.’

However, there was one instance in the MBC in example (11) which shows the preposition *ki* functioning as the phrase type marker of the cause phrase *ki te wai Māori*. This is the only example in the MBC like this. All other adverbials in this type of environment in the MBC have *i* as the phrase-type marker.

11.	Kī	tonu	te	rākau	ki	te	wai	māori
	full	still	theSG	tree	with	theSG	water	māori
	‘The tree is still full of the drinking water’							

A minimal state intransitive sentence consists of a verb constituent and a subject. Where there is no adverbial expressing cause and only the subject noun phrase, the collocates of *kī* ‘full’ were useful indicators of what lexeme was the lexical head of the phrase. We see in example (12) *ngā tūru* ‘chairs’ could only sensibly function as the subject phrase to *kī* ‘full’, where it is the patient, and not as the subject noun phrase of *kī* ‘say’, where it is the agent. In example (13) the collocate *kete* regularly co-occurs with the sense ‘full’. The collocates of these senses will be further developed in Table 5.10.

12.	Kī	katoa	ngā	tūru
	full	all	thePL	chair
	‘The chairs were all full.’			

13.	Ka	kī	te	kete
	TAM	full	theSG	bag
	‘The bag will be full.’			

As mentioned in the grammar review, action intransitives where the lexical head is *kī* ‘say’ sometimes include an adverbial expressing the goal, as in the *ki* phrase in (14) and example (15) where the particle *mai* indicates the goal, and the direct object is ‘what is said’ and takes the form of a sentence, and is in inverted commas.

14.	Ka	kī	atu	ia	ki	a	rātou
	TAM	say	away	IIISG	to	pers	IIIPL
	‘He spoke to them.’						

15.	Ka	kī	mai	ia	“ka	hui	mātou
	TAM	say	hither	IIISG	TAM	meet	Iplexcl
	ā	tēnei	pō”				
	atFUT	this	night				
	‘He said to me, “We will meet tonight.”’						

In cases where the lexeme *kī* ‘say’ occurred in an action intransitive sentence, the goal phrase which followed exemplified by *ki a ia* in (16) was fundamental to

recognising this lexeme. There were ambiguous examples in this environment where there was no goal phrase such as example (17), where the sense of *kī* could be understood as either ‘say’ or ‘full’. Because the ambiguity lies in the subject NP it could be understood as a patient of a state intransitive sentence with the sense of ‘full’ or the actor of the sense ‘say’. In the majority, the post-head modifying particle such as *atu* in (16) and *mai* in (15) above identified it as the sense of ‘say’. However there were also examples of *kī* ‘full’ with post-head modifying particles. The post-head modifying particles are looked at further in Table 5.8.

16. Kī atu au ki a ia
say away ISG to pers IIISG
‘I spoke to him.’

17. E kī ana ngā ripoata nei
TAM say TAM thePL report here
‘These reports are full.’ ‘These reports spoke.’

4.3.2 Phrase type markers – TAMs

There are certain parts which make up a phrase which can tell us more about a word’s sense. For example, Bauer (1997:8) states that the lexical head of a phrase with a TAM marker is a verb. We can start by looking at the phrase type marker which precedes *kī.* This process produces all verbal forms of *kī* excluding examples containing the TAMs *i te, ki te* and *kei te* that precede *kī.* These last three TAMs cause ambiguity due to the orthographic convention which writes the determiner as a separate word e.g. *kei te*. A manual search and coding of these TAMs however has sorted them into their correct environments. There are issues surrounding stem nominalisations which can be found following the determiner *te* which are nominal in form yet verbal in sense. These issues will be discussed shortly.

The first step is to consider syntactic criteria such as co-occurring phrase type markers which indicate a verbal use of kī or a nominal use. So for example where the phrase type marker *ka* occurs preceding *kī* as in example (18) we can automatically assign that to a verbal sense, eliminating the nominal ‘key’ sense.

18.	ka	kī	mai	ia	ki	a	mātou
	TAM	say	hither	IIISG	to	pers	IPLEXCL

‘he said to us…’

Though there are examples from other sources which exemplified the use of the lexeme *kī* ‘key’ as the lexical head of a verbal predicate as in (19), there was only one such example found in the MBC. This is fairly uncommon, but I was able to search further using Think Tank. (Think Tank is a computerised search tool created for Te Taura Whiri i te Reo Māori for use with their extensive Matapuna corpus of Māori language materials. It is not available for general use, but access was granted for a small number of language research students at Te Kawa a Māui, Victoria University of Wellington. I would like to acknowledge my gratitude to Te Taura Whiri for their generosity in making this resource available to me.) This search produced only two results, example (19) being one of them.

19.	e	kī	ana	te	kēti	ki	te	kī
	TAM	key	TAM	theSG	gate	with	theSG	key

‘the gate was locked with the key’

If we look at the results from the data about *kī* we notice that not only do TAMs distinguish verbal uses from nominal uses, they also give clues that distinguish verbal from verbal uses, in this case *kī* ‘say’ from *kī* ‘full’. We will look firstly at the results of the individual TAMs preceding *kī*. The following table shows a breakdown of the TAMs which co-occurred with the lexeme *kī* ‘say’ from most frequent to least. The TAMs that preceded *kī* ‘full’ have also been given for comparison.

Table 4.6 Tense Aspect Mood Markers which co-occurred with *kī*

TAM	'say'	'full'	2-tailed P	X2	sig
me	552	0	0.0002	13.602	extremely
ka	221	8	0.2806	1.164	not
e … ana	219	4	0.5525	0.353	not
e	177	2	0.2612	1.262	not
i	93	2	0.8497	0.036	not
kei te	76	2	0.9253	0.009	not
ø	71	13	>0.0001	61.296	extremely
kua	50	5	0.0012	10.504	very
kia	9	2	0.0006	11.693	extremely
i te	7	0	0.6779	0.172	not
kei	2	1	0.0005	112.253	extremely
Total	1477	39			

For *kia*, *i te* and *kei*, the numbers of tokens are too small to place any reliance on the results.

Interestingly *me* is the most frequent TAM preceding *kī*, and it occurs only with the 'say' lexeme. The nature of the spoken corpus is a probable explanation for its high frequency. The majority of instances were pauses in speech while the speaker thinks about the word needed as in (20).

20. ko ngā me kī ko ngā ture
PREP thePL TAM say PREP thePL law
'the … 'it should be said' … the law'

It is difficult to verify this hypothesis from the transcripts, though it could be done by listening to the recordings to determine whether or not the speaker pauses on *me kī,* but from looking at the environments in which it occurs it seems likely that this is, in fact, what is happening.

Me kī may also be used by the speaker to signify that some relevant piece of information is about to be said (21), or to elaborate on information already stated (22). *Me kī* also seems to be used for emphasis, as in (23).

21. Me kī rā nā koutou i tangi
TAM say DIST belong IIIPL TAM cry,
ā, nā tātou katoa i tangi
and, belong IPLINCL all TAM cry
'[It] should be mentioned that when you cried, we all cried [with you].'

22. Ahakoa kei reira tonu te kai rā,
although atPRES there cont theSG food away,
me kī he kōura, he aha rānei...
TAM say a crayfish a what or
'Although there is still food there, that is, crayfish and whatever else …'

23. ...ko te wahine, ā, koinā, me
…EQ theSG woman uh that is TAM
kī ake te whare tangata
say up theSG house person
'…the woman, uh, that is, [she] should be called the child bearer'

Thus there were many senses of *me kī* in the MBC. Examples 21-23 represent those uses which are highly likely to occur only in spoken data such as that in the MBC. It suggests that the use of *me kī* 'one should mention' or 'it should be said' is in regular use in spoken data as this function would not necessarily be useful in written text. In written text, you do not need to fill your thinking pauses. The TAM *me* never occurred preceding *kī* 'full' in the corpus. There are structural constraints on the use of a state intransitive verb as an imperative, or

in obligation expressions – one cannot “oblige” an inanimate object, and therefore the sense *kī* ‘full’ could never co-occur with *me*. Therefore, the results of the data in the MBC on *kī* and the structural constraints on *kī* ‘full’ as a state intransitive verb shows that a tagger could be 100% certain that where *kī* co-occurs with *me* the lexeme can be tagged as the sense ‘say’.

The results show that where the TAM *ka* co-occurs with *kī*, it has a 96.5% probability of being the 'say' sense, and a 3.5% probability of being the 'full' sense. These percentages are very similar to the percentages for these lexemes in the MBC overall: 97.5% of the tokens of *kī* are 'say', and 2.4% are 'full'. Thus the results of the two-tailed P values test for *ka* are, as expected, not significant. Where the sense of *kī* ‘full’ co-occurred with *ka* it was the nature of the subject, and any adverbial expressing cause, such as *i te kai* in (24), *i ngā whare* in (25) and *i te pāua* in (26) that signalled the right interpretation of *kī* ‘full’.

24. Ka kī te kāpata i te kai
TAM full theSG cupboard cause theSG food
‘The cupboard is full of food.’

25. Ka kī katoa te whenua i ngā whare
TAM full all theSG land cause thePL house
‘The entire land is full of houses.’

26. Ka kī tō kete i te paua
TAM full yourSG bag cause theSG paua
‘Your bag is full of paua.’

The continuous TAM *e … ana* is the third most frequent TAM to co-occur with the lexeme *kī* ‘say’. Of these examples, 98% co-occurred with ‘say’ and 2% co-occurred with ‘full’. As with *ka* these are quite similar to the overall percentages, and so the two-tailed P value is not significant.

The TAM *e* was the fourth most frequent TAM occurring before *kī.* The TAM *e* can function as the phrase-type marker of the multi-word unit ***e kī*** which is recognised as a multi-word unit with its own sense by all the dictionaries

consulted. The sense that was associated with *e kī* was ‘you don’t say’/‘gosh’. Only 15 examples can be attributed to the multi-word unit *e kī*, as in (27). These uses were found co-occurring with an exclamation mark as in (27) or the phrase was repeated as in (28):

27. E kī! He mea kino Pāpā
TAM say cls thing bad dad
‘Gosh! That’s bad, Dad.’

28. E kī e kī rā Hine-wehi!
TAM say TAM say away Hine-wehi
‘Oh my goodness, Hine-wehi!’

It is worth mentioning that there were two examples where *e kī* occurred with the post-head modifying particle *rā* as in (28).

The remainder of the examples of *e kī* ‘say’ function as the verb constituent of subordinate clauses as in (29) and (30).

29. Kāore mātou e kī atu
NEG IPLEXCL TAM say away
‘We didn’t say anything.’

30. Ka rongo ake i te rangatira e kī
TAM hear EMPH DO theSG chief TAM say
nei
here
‘’[You] will hear the chief who is speaking here.’

The majority of the examples where the TAM *e* preceded *kī* ‘say’ were subordinate clauses of negative sentences as in (29), and the verb constituent of the actor-emphatic construction as in (31).

31. Māku e kī atu
belong ISG TAM say away
‘I will say…’

The two instances where *e* preceded the lexeme *kī* ‘full’ were in the subordinate clause of a negative sentence as in (32) and (33):

32. Kua kore e kī ngā puku o
TAM NEG TAM full thePL stomach of
ēnei tamariki
these children
'The stomachs of these children are never full.'

33. Kāore e kī tō puku
NEG TAM full yourSG stomach
'Your stomach isn't full.'

The data from the MBC on *kī* shows that where the word *kī* co-occurs with *e* outside of the environment of negative clauses it was the sense 'say'. Where the sense 'full' occurs in the two negative clauses in (32) and (33) it was the collocate *puku* in the subject phrase that was the distinguishing factor between senses.

The results for the TAM *i* showed the two senses are very like the overall proportions with 98% of uses occurring with the 'say' sense and 2% of uses occurring with the 'full' sense. Therefore, *i* does not provide any tagging predictions.

The overall percentages are again similar for *kei te* co-occurring with the lexeme 'say' 97% of the time and 'full' 3% of the time. Therefore the two-tailed P value is not significant.

There were 71 instances where there was Ø TAM in the verb constituent that co-occurred with the lexeme *kī* 'say' as the lexical head and 13 instances of *kī* 'full' in this environment. Statistically, Table 4.6 shows this environment is significant for signalling the lexeme *kī* 'full', suggesting that *kī* 'full' is more likely to occur in the environment of Ø TAM than the 'say' sense. The two-tailed P value was extremely significant for this TAM. All examples of *kī* that occurred in the verb constituent with Ø TAM co-occurred with post-posed modifiers in the phrase. The most prevalent modifier was *tonu* of which there were 7 examples following the 'full' sense and only 2 examples following the 'say' sense.

The sense 'full' was three times more likely to co-occur with the TAM *kua* than the sense 'say'. *Kua* occurred preceding the 'say' sense a total of 90% and the 'full' sense a total of 10% which according to the two-tailed P value is very

statistically significant.

The TAM *kia* had an even distribution between senses with the sense ‘say’ occurring 81% and ‘full’ 19%. Though the two-tailed P value test showed this distribution to be extremely significant in favour of *kī* 'full' the number of tokens is so small that it would be unwise to make predictions on this basis.

All instances of the TAM *i te* were attributed to the ‘say’ sense. All examples showed the TAM *i te* functioning as the phrase type marker of the subordinate clause of a negative sentence.

There were only three examples of the TAM *kei* as the phrase-type marker of the verbal predicate. Two examples were attributed to the sense ‘say’ and one example to ‘full’. *Kei* was shown as statistically significant for signalling the lexeme *kī* ‘full’. The data set for this environment however was only very small, accounting for only three cases of *kei* co-occuring with ‘full’, so is not a trustable result for tagging purposes.

4.3.3 Determiners

The next step was to look at what the environment of determiners as the phrase-type marker might suggest regarding the sense *kī* ‘full’ and *kī* ‘say’. In all cases for the sense ‘say’ and ‘full’ the phrase in which it occurred was a stem nominalisation. The most frequent determiner to co-occur with *kī* was *te*. There was only one instance of *kī* ‘full’ in this environment, one instance of ‘key’ (the only example of ‘key’ in the entire corpus) and 89 instances of *kī* ‘say’.

Table 4.7 Determiners

Det	**‘say’**	**‘full’**	**‘key’**
te	89	1	1
tana	8		
he	7	1	
taku	5		
tā rātou	3		
tā mātou	2		
tāna	2		
tō	2		
tā tāua nei	1		
Total	**119**	**2**	**1**

The P value calculations would not reveal anything useful here, because of the rarity of determiners with the ‘full’ and ‘key’ senses.

The determiner *te* ‘the’ was most commonly used, as is expected, for nominalisations. As previously mentioned, stem nominalisations cause issues when tagging for this type of environment. However, the low frequency of the sense ‘key’ minimises the confusion, so that counts of determiners preceding *kī* can be assigned to the ‘full’ and ‘say’ sense. In the two cases where ‘full’ was preceded by a determiner, it was the syntax which pointed to the correct lexeme: the cause phrase *i te waipiro* in (34) and *i te moni* in (35), and the collocate of *tōku pēke* also in (35) signalled the correct sense.

34. He kī nei i te waipiro
    CLS full here cause theSG beer
    ‘[It] was full of beer.’

35. Kore kē he painga te kī o tōku
NEG CONTR CLS good theSG full of my
pēke i te moni
bag cause theSG money
'What good is pockets full of money?'

A look into those modifiers that co-occurred with stem nominalisations did not produce any significant results. There were 56 examples out of 132 stem nominalisations with modifiers. The clause position did not signal a nominal or verbal sense of *kī* in this environment either.

4.3.4 Modifiers

Despite the situation with stem nominalisations, the modifier in the phrase can sometimes help to differentiate between senses. For example, the results in Table 4.8 show that the modifying particle *mai* only occurred in the phrase following the sense 'say'. This is a significant finding, in that *mai* was also the most frequent particle to occur in the phrase. There were a total of 209 uses of *mai* of 703 various particles in the phrase. Although the directional particles *mai* 'hither' and *atu* 'away' would semantically be expected to co-occur only with *kī* 'say', there were two instances of the directional particle *atu* in a phrase with the sense 'full', see examples (36) and (37). It is interesting to note that the commonality between the two examples is the Ø marking of the TAM. Comparatively, 98.9% of the instances of *atu* were found in a phrase with the sense 'say' 1.5% in the phrase with the sense 'full'. Bauer (1997:350) states that directional particles are not necessarily used only to mark physical movement but also to mark mental orientation which could explain why *atu* co-occurs with 'full'. The fact that we are dealing with spoken data affects the results, as Bauer also states that *mai* is far more frequent in first-person discourse. This could affect the validity of *mai* as a signal for the sense 'say' in a written corpus because of the special way it behaves in spoken discourse.

36. Kī atu i te kapū tī me te kai
full away cause theSG cup tea with theSG food
'[It] was full of the cup of tea and the food.'

37.	Kī	atu	i	tēnā	mea	i	te,	i
	full	away	PREP	that	thing	cause	the	cause
	te	kuku						
	theSG	mussel						

'[It] was full of, of mussels.'

There were a total of 703 modifying particles that co-occurred with the sense 'say' and 22 modifying particles co-occurred with the sense 'full'. Statistically 97% of those examples of *kī* which co-occurred with modifying particles in the phrase were the 'say' sense and 3% were the 'full' sense.

Table 4.8 Modifiers to *kī*

Modifier	'say'	'full'	X2	2-tailed P	sig
MAI	209		7.580	0.0059	very
ATU	189	2	3.402	0.0651	not quite
RĀ	64		2.321	0.1276	not
AKE	55		1.995	0.1578	not
NEI	40	2	0.198	0.6563	not
AI	29	2	0.800	0.3712	not
ANŌ	22		0.798	0.3717	not
PEA	20		0.725	0.3944	not
HOKI	9	1	1.251	0.2634	not
PAI	5	1	3.080	0.0793	not quite
TONU	5	9	153.157	>0.0001	extremely
NOA	3		0.109	0.7415	not
KATOA	2	4	70.881	>0.0001	extremely
KĒ	2		0.073	0.7877	not
NĀ	1		0.036	0.8490	not
RAWA	1	1	12.804	0.0003	extremely
Total	**703**	**22**			

Mai, *tonu*, *katoa* and *rawa* produced two-tailed P values that were significant, but the numbers for *rawa* and probably *katoa* as well are too small to place much reliance on them.

The most telling result here is the extremely significant P value for *tonu*, which occurs 64% of the time with the 'full' sense of *kī*, and only 36% of the time with the 'say' sense. However, despite the P value, these percentages warn that it would be rash to tag *tonu* automatically as the 'full' sense. However, there was Ø marking of the TAM in seven of the nine examples of *tonu* co-occurring with 'full', and these two features together could reliably be tagged for 'full'. The 'very significant' result for *mai* is probably also helpful for tagging purposes, as indicative of the 'say' sense.

4.3.5 Sentence/Clause position of 'full' and 'say'

Other syntactic elements were also examined for each example of *kī* that is, sentence/clause position and other contextual clues such as key words or collocates which co-occur with *kī* such as the modifier in the phrase. Where *kī* functions as the lexical head of a verb phrase any causer phrase and goal phrase of *kī* (when present) has also been noted.

The sentence/clause position can shed further light on the types of environments in which we find the senses of *kī* in Māori. In this case study of *kī* the categories of syntactic criteria involved in the corpus data analysis are: the sentence/ clause position whereby examples were labelled with either 'Prepositional Phrase' (PP) where *kī* functioned as the lexical head of a prepositional phrase as in example (38), Predicate Head (PH) shown in example (39) where *kī* is the lexical head of an equative nominal predicate, and example (40) where *kī* functions as the lexical head of a verbal predicate. The phrase type markers were noted for each use of *kī* in the environments of 'PH' which identified those examples as being either verbal or nominal.

38. me tana kī atu anō
    with hisSG say away again
    'with this he said…'

39.	ko	te	kī	o	te	kūaha	tēnei
	PREP	theSG	key	of	theSG	door	this
	‘this is the key to the door’						
40.	ka	kī	atu	te	rangatahi		nei
	TAM	say	away	theSG	youth		here
	‘and this youth said…’						

Where *kī* functioned as the lexical head of a subject noun phrase it was marked (SNP) as in example (41):

41.	Kotahi	rau	paiheneti	tā mātou	kī	atu
	one	hundred	percent	ourSGEXCL	say	away
	ki	a	koutou	ināianei		
	to	PERS	IIPL	now		
	‘They are offering 100% support’					

The table below presents the results from the data in the MBC. The overall picture we can draw from this is that the sense ‘full’ never occurred in a prepositional phrase, *taea* complement or actor-emphatic clause, yet the sense ‘say’ did.

Table 4.9 Comparison of the sentence/clause position of *kī*

Sentence/clause position	**'say'**	**'full'**	**'key'**	**X2**	**2-tailed P**	**sig**
Verbal Predicate head	1406	36	0	0.4299	0.0381	sig
Prepositional Phrase	60	0	0			
Nominal Predicate Head	30	1	0			
Neg VC	29	2	0			
Agent Emphatic Verb Clause	26	0	0			
Subject Noun Phrase	25	1	1			
Taea complement	17	0	0			
Total	**1593**	**40**	**1**			

Because there is only one token of 'key' in the MBC, it has been ignored for statistical purposes. The low numbers of 'full' in the other environments render statistical calculations unhelpful.

It comes as no surprise that the most frequent environment of both senses was found to be the verbal predicate. Although the P value for Verbal Predicate head was significant, this environment is not helpful for tagging the difference between 'full' and 'say', since that is the usual environment for both lexemes. The second most frequent environment is the prepositional phrase which the lexeme *kī* 'say' could be tagged for. There were no occurrences of *kī* 'full' in the prepositional phrase. The next most frequent environment was the nominal predicate: only one 'full' sense occurred there, in contrast to 30 senses of 'say'. All occurrences in the nominal predicate were stem nominalisations.

4.3.6 Semantics and Context

The most common distinguishing factor between the lexeme 'say' and the lexeme 'full' was often expressed in the subject noun phrase and the adverbial phrase expressing cause of the lexeme 'full'. Where there was a subject noun

phrase and an adverbial expressing cause it was clear it was the ‘full’ sense because of the matching of the collocates of ‘full’ such as *te wai* ‘the water’ and *i ngā tuna* ‘with eels’ in (42). These types of collocates, that is, inanimate subject noun phrases and vessels or containers of sorts, were key to distinguishing between the two lexemes.

42.	E	kī	ana	te	wai	nei	i	ngā	tuna
	TAM	full	TAM	theSG	water	here	by	thePL	eels
	‘The water is full of eels.’								

43.	E	kī	ana	i	te	wai	hopi
	TAM	full	TAM	by	theSG	water	soap
	‘[It] was full of soapy water.’						

44.	E	kī	ana	i	te	waka
	TAM	full	TAM	by	theSG	boat
	‘[It] was full of boats.’					

45.	E	kī	ana	i	te	hūpē
	TAM	full	TAM	by	theSG	snot
	‘[It] was full of snot.’					

The collocates expressed in (43) – (45) are items which are important contextual clues for identifying the lexeme ‘full’ in a corpus. These collocates - ‘soapy water’, ‘boats’, ‘snot’ are types of words which signal the lexeme ‘full’ and not ‘say’ and are clear contextual clues for identifying this lexeme.

All examples listed in Table 4.10 are collocates that signalled the lexeme ‘full’. As previously mentioned the collocates of the sense ‘full’ that occurred in the subject NP were very different to those collocates that occurred in the subject NP where the lexeme *kī* ‘say’ functioned as the lexical head of the VC. The collocates in the subject NP in Table 4.10 are inanimate ‘vessels’ that were key indicators of the lexeme ‘full’. The adverbial expressing cause was also a very significant indicator of the ‘full’ sense. The preposition *i* ‘by’ marks an adverbial expressing cause and was a key factor signalling the correct interpretation of ‘full’. As mentioned in the grammar review, the action intransitive use of *kī* ‘say’ could only ever have the preposition *ki* as the phrase-type marker of the goal phrase. The canonical transitive verb can have any type

of phrase as the DO yet in this particular environment, the subject NP could signal the right interpretation as animacy of the subject noun phrase contrasted with *kī* 'full' and signalled the correct interpretation of *kī* 'say'.

The collocates in Table 4.10 represent a total of 38 of the 41 examples extracted from the MBC. The Table illustrates the patterns of the semantic connections between the subject noun phrases of *kī* 'full'. It also shows those collocates in the adverbial expressing cause and how these relate to 'full'.

Table 4.10 Collocates of *kī* 'full' in the subject NP and Adverbial expressing cause

Subject NP	**Adverbial expressing cause**
te kāpata 'the cupboard'	i te kai 'by the food'
ngā puku o ēnei tamariki 'the stomachs of these children'	
te kete 'the basket'	
te whenua 'the land'	i ngā whare 'by the houses'
tō mātou whare 'our house'	
tērā whare 'that house'	i ngā pounamu 'by the greenstone'
tō kete 'your basket'	i te paua 'by the paua'
te kete 'the basket'	
te whare 'the house'	i te kaumātua 'by the elders'
Topicalised ko ēnei ipu 'these containers'	i te wai hopi 'by the soapy water'
tēnei rākau 'this tree'	ki te wai Māori 'with the drinking water'
[understood]	o tōku pēke i te moni 'of my bag by the money'
te kete 'the basket'	
te whare 'the house'	i ngā peira 'by the bailer'
	i te tangata 'by the person'
te pēke 'the bag'	
[understood]	i te kuku 'by the mussels'
tō peke 'your bag'	i te pikopiko 'by the pikopiko'
ngā papa o te whare nei 'the floors of this house'	i ana tāonga 'by his treasures'
ēnei nā 'these here'	i te tēneti 'by the tents'

Subject NP	**Adverbial expressing cause**
te onepū ‘the sand’	
ērā kete kōrero ‘those conversational baskets’	
Topicalised tēnei moana a Pēwhairangi ‘this ocean of Pēwhairangi’	i te waka ‘by the boats’
te kete ‘the basket’	
tō tātou whare ‘our house’	
Topicalised tōna rūma ‘his room’	i ngā taonga katoa ‘by all of the treasures’
te awa nei ‘this river’	i ngā tuna ‘by the tuna’
te motu nei ‘this land’	i te haunga o te ika ‘by the smell of the fish’
[understood]	i te moni ‘by the money’
te kāpata ‘the cupboard’	
te pataka ‘the library’	
[understood]	i te hūpē ‘by the snot’
tō puku ‘your stomach’	
[understood]	i te kapū tī me ngā kai ‘by the cups of tea and food’
te puku o te tamaiti ‘the stomach of the child’	
ngā tūru ‘the chair’	
te mihini nei ‘this machine’	
te takere o te waka ‘the hull of the boat’	i te ika mōmona nei ‘by these fatty fish’

All the collocates in Table 4.10 indicate the sense ‘full’. The main contrast here is the semantic content of the subject NP for each sense. The ‘full’ sense contains vessels and inanimate entities in the subject NP in comparison to animate beings that occur in the subject NP of the ‘say’ sense.

4.4. Conclusions

The information gathered from both the section on etymology and the section on the analysis of *kī* tells us that we have three senses of *kī*: ‘say’, ‘full’ and ‘key’. Due to the infrequent use of ‘key’ in the corpus, the analysis focused only on ‘full’ and ‘say’. The most significant factor that distinguished the sense ‘full’ from ‘say’ was the collocates that occurred in the subject NP. Where there were ambiguous examples such as ‘ahau’ in the subject NP, ambiguous because the animate being *ahau* ‘I/me’ could be ‘full’ of *kai* ‘food’ for example, it was the adverbial expressing cause which signalled the ‘full’ sense (where there was an adverbial available to make this distinction). The presence of a goal phrase marked by *kī* signalled the ‘say’ sense. There were other forms which could confidently be tagged for one lexeme or the other, such as the phrase-type marker *me.* The results concluded that *me* co-occurred only with the ‘say’ sense, though the multi-word unit *me kī* could be restricted to a spoken corpus due to its function. Its use as a pause phrase could explain its high frequency in the MBC. The next most significant phrase type marker to co-occur with the sense ‘say’ was *e*. There were only two instances where *e* co-occurred with ‘full’ and both examples functioned as the lexical head of the subordinate clause in a negative sentence. Therefore *e* could be tagged for the ‘say’ sense except in the environment of a negative sentence where other factors would have to be considered. Statistically where there was *Ø* TAM marker in the verb clause it was highly likely to co-occur with the ‘full’ sense. Comparatively the ‘full’ sense was three times more likely to co-occur with the TAM *kua*. The TAM *i te* only ever co-occurred with the ‘say’ sense, functioning as the TAM marker of the subordinate clause in a negative sentence. The results of determiners as the phrase-type markers show a high occurrence of *te* preceding *kī* with the ‘say’ sense, these results were not significant using two-tailed P values, but this could be to do with the small sample set. There was only one example of ‘full’ to follow *te* and one example of ‘key’. There was one example which showed the determiner *he* co-occurring with the ‘full’ sense and seven examples with the ‘say’ sense. The modifier *mai* only ever co-occurred with ‘say’ and there was a high frequency of *tonu* co-occurring with ‘full’. Statistically 97% of those examples of *kī* which co-occurred with modifying particles in the phrase were

the ‘say’ sense and 3% were the ‘full’ sense. The sentence-clause position showed us that ‘full’ never occurred in a prepositional phrase, *taea* complement or actor emphatic clause. No individual syntactic position was indicative of one sense rather than the other. There was only one occurrence of ‘full’ as a nominal predicate head. The results from the data on *kī* show that it is possible to discriminate between *kī* ‘full’ and *kī* ‘say’ in a corpus by using both syntactic criteria and items in the phrase periphery.

5 Mea

5.0 Introduction

This chapter looks at *mea* using the same methodology as for *kī*. The same dictionaries and etymological resources provided the evidence for determining the appropriate lexemes for *mea*. The Māori Broadcast Corpus (MBC) has again provided the data for *mea* and has been manually tagged for its various senses for the purpose of this research. This step provides a reliable diagnostic for analysis in the corpus. The results section provides information about the data in the MBC and looks at the frequency of each lexeme and the environments in which it occurs. The final analysis and results section of this chapter details how a tagger might tag for each sense of *mea*.

5.1 Establishing lexemes associated with *mea*

5.1.1. Dictionary analysis of mea

The following is a selection of information from the reputable dictionaries mentioned in the first section of this study. The information has been placed into sections in Table 5.1. Each column is headed with the dictionary name, and the information from each dictionary has been placed in the following order: meaning – the meanings are entered into the table as they are recognised in the dictionaries, so if a word or words has been classified under the same head entry, they will be entered under the same meaning; associated grammatical functions, and finally, senses (senses have been entered according to the senses the dictionaries have associated with the meaning in the head entry). The associated grammatical functions are abbreviated as follows: Noun (N), Indefinite Pronoun (IPN), Transitive Verb (VT), State Intransitive Verb (VS), Action Intransitive Verb (VI).

Table 5.1. Information from dictionaries about *mea*

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Tirohia Kimihia***	***He Pātaka Kupu***
mea¹ *Thing/say*	*√*	*√*	*√*	*√*	*√*
Associated grammatical functions:	*N, IPN, VT*	*N, IPN, VT*	*N, IPN, VT*	*N, IPN, VT*	*N, VT, VI, VS*
Senses:	*Thing, so-and-so, what's his name?* *Say*	*Thing, object, one, reason, thingy, the one, that thing* *Say, speak, do, deal with, think, intend, make, use*	*Thing, reason/cause, fact/event/case, one, so-and-so* *Say, intend/wish, think*	*Thing* *say, think, wish, so-and-so*	*Thing* *Instruct, wish*
mea² *Red/reddish*			*√*		*√*
Associated grammatical functions:			*VS*		*VS, VI, N*
Senses:			*Red/reddish*		*Red/reddish*
mea³ *Mayor*		*√*			
Associated grammatical functions:		*N*			
Senses:		*Mayor*			

In Table 5.1 Williams and *HPK* both recognise two lexemes associated with the word-form *mea*. The meanings attached to those two lexemes are ‘thing/say’, and ‘red/reddish’. Moorfield also distinguishes between two separate lexemes; however, the meanings associated with these lexemes are ‘thing/say’ and ‘mayor’. Biggs lists two entries for *mea,* first is ‘thing/say’ and then ‘so-and-so’.

According to the dictionaries reviewed here we have three likely lexemes associated with *mea* in Māori, that is, ‘thing/say’, ‘red/reddish’ and ‘mayor’. It is interesting to note that none of the dictionaries has separated the senses of *mea* ‘thing/say’ but instead all consider them as one lexeme with these two very different senses. The reason for the lack of divide between the two senses ‘thing’ and ‘say’ is not clear. This will be discussed further in section 5.3.

The grammatical functions given by these various sources are as follows: the sense ‘thing’ has been classed as a noun and an indefinite pronoun by all dictionaries except Moorfield. The sense ‘say’ has been classed as a canonical transitive verb by all five dictionaries. *HPK* has the grammatical labels canonical transitive verb, action intransitive verb, noun and state intransitive for the ‘say’ sense. *HPK* also classes the sense ‘red/reddish’ as a state intransitive verb, an action intransitive verb and a noun. *Williams* also identifies *mea* ‘red/reddish’ as an adjective which is equivalent to what I am calling a state intransitive.

5.1.2. Etymology

Tregear (1891) has one entry only for the word *mea,* listed in Table 2.2 below, whereas Greenhill & Clark (2011) has two.

Table 5.2. Senses of *mea* listed by Tregear (1891) and Greenhill & Clark (2011)

Tregear		*Greenhill & Clark (2011)*	
	Thing: *a. A word used as a subsitute for another noun* *b. such an one (Mr What's-his-name)* *c. To do* *d. To cause* *e. To say* *f. To intend, to wish* *g. To think* *h. A lapse of time* *i. A thing of no consequence*		*Thing:* *a. (Any)thing;* *b. do (anything)* *c. what's-his-name*
			Reddish

The senses of *mea* in Greenhill & Clark (2011) in Table 5.2 are separate lexemes which are related to the lexemes in other Polynesian languages as shown in Table 5.3. The results of the analysis of the data from the MBC did not produce any examples containing the sense 'red/reddish'. This sense has been excluded from the Polynesian comparative review due to the lack of use in the MBC.

Table 5.3 below is a comparative account of the Polynesian lexemes considered to be related to the lexeme *mea* 'thing/say' in Māori. The dictionary review did not provide evidence of *mea* 'thing' and *mea* 'say' as being distinct lexemes; however as a speaker of Māori, I feel that there is a semantic difference between these two senses, and therefore it is worth looking at how these senses are related throughout Polynesia. Justification to support my claim that we are dealing with two distinct lexemes of *mea* is given at the end of this section.

Table 5.3. Polynesian lexemes and meanings associated with *mea* 'thing/say' listed by Tregear (1891) and Greenhill & Clark (2011)

	Tregear		***Greenhill & Clark (2011)***	
PN Language	***Lexeme***	***Sense***	***Lexeme***	***Sense***
Samoan	*mea*	*a thing; a place; an animal or live creature; a creature, applied to persons; the private parts, when used idiomatically; to do, to prepare.*	*mea*	*thing*
Tahitian	*mea*	*a thing, a person; anything previously mentioned; so-and-so; to do, a word used as a convenient substitute, instead of naming the action.*	*mea*	*a thing, a person, anything mentioned*
Hawaiian		*a thing, an external object; a circumstance or condition; a person, a thing, in its most extensive sense; Having the quality of obtaining or possessing something; to do, to say, to act; to meddle with; to touch, to injure; to trouble with unprofitable business; to hinder; to cause to come to; to speak, to utter; to ask questions*	*mea*	*thing*

	Tregear		***Greenhill & Clark (2011)***	
Tongan	*mea*	*things in general; matters; property; affairs; to do; to look at, to attend to. Cf. meai, to know, to be acquainted with; femeaaki, to converse (applied to chiefs).*	*meʔa*	*thing*
Rarotongan	*mea*	*a thing*	*mea*	*thing*
Marquesean	*mea*	*a thing; an individual; to do; to do a bad action; meamea, a joke; pleasantry.*	*mea*	*thing*
Mangarevan	*mea*	*a thing*	*mea*	*a thing*
Pukapuka				*thing, say, think, do*
Luangiua			*mea*	*thing* *reply*

The lexemes in Table 5.3 provide information as to what languages throughout Polynesia recognise the sense ‘thing’ and those that recognise the sense ‘say’. All the Polynesian languages listed in Table 5.3 recognise the sense ‘thing’. It is interesting to note that there has not been any change in form of the word *mea*. Greenhill & Clark (2011) lists the word form *me?a* though Tregear does not. Only some of the languages contain cognates of ‘say’. Tregear lists the sense ‘say’ in Hawaiian though no equivalent data is provided by Greenhill & Clark (2011). Tregear also lists the word form *fe**mea**aki* ‘to converse’ in Tongan. The Polynesian languages provided by Greenhill & Clark (2011) with the sense ‘say’ are: Pukapuka ‘to say, think, do’ and Luangiua which has ‘reply’ as a sense. It is interesting to note that the entry for Luangiua found in Greenhill & Clark (2011) recognises three distinct lexemes of *mea*, ‘thing’, ‘reply’ and ‘red/brightly coloured’. The evidence suggests that there are randomly dispersed remnants of *mea* ‘say’ throughout Polynesia, from Eastern Polynesian languages MAO, Pukapukan and Hawaiian, to the Samoic-Outlier language Luangiua and finally in an early branch of the Polynesian family tree, in the Tongic language, Tongan.

There are no lexemes cognate with the Māori word *mea* ‘mayor’ in this etymological information. This is because *mea* ‘mayor’ is a loan word in Te Reo and because these sources are comparing indigenous Polynesian words. There are only ten examples of *mea* ‘mayor’ in the MBC.

5.1.3. Lexeme summary

Overall, the dictionary review and etymology review did not support *mea* ‘thing’ and ‘say’ as being two distinct lexemes, with the possible exception of Luangiua, where ‘thing’ and ‘say’ appear to be distinct lexemes (see bottom of Table 5.3). This is the only evidence to support my claim that these senses are now considered distinct. There was very little evidence of *mea* ‘say’ as being a distinct lexeme in other languages of Polynesia. It is quite possible that because the word form *mea* has not changed form for either sense there is no evidence to support these senses being classed as distinct and perhaps there is reason not stipulated in the literature for the sense ‘say’ and ‘thing’ to be etymologically related.

In an attempt to support my claim that *mea* ‘thing’ and ‘say’ are in fact recognised as distinct lexemes by today’s speakers of Te Reo, I asked ten native/proficient speakers of Te Reo to look at two sentences in Māori. The sentences contained the ‘thing’ sense in one and the ‘say’ sense in another. When asked if they felt that the senses ‘say’ and ‘thing’ used in the Māori sentences were related, all ten informants answered that no they were not, they considered them distinct lexemes with very different meanings. The fact that this is not reflected in the dictionaries consulted can probably be explained by the fact that they have copied each other, and that the Williams dictionary based its entry on the etymological information.

If there are two distinct lexemes, and they are of Polynesian origin (as we assume), then we would expect to find traces of both in other Polynesian languages. Since we do not find this, we are led to the belief that the sense ‘say’ was a Maori innovation. The most likely source of this innovation is that it was an extension of sense from the original lexeme ‘thing’. ‘Thing’ and ‘say’ would thus historically be different senses of one polysemous lexeme, ‘thing’. Over time, as the ‘say’ sense became established, speakers developed the intuition that these senses had no relationship, and thus now regard these two as separate homonyms. Since Lyons (1977:550) regards native speaker intuition as an important criterion for distinguishing polysemes and homonyms, I conclude that there is a case to be made for regarding these in Maori as separate lexemes

The process is likely to have its roots in the fact that *mea* ‘to do’ is used as a general filler verb, just as *mea* ‘thing’ is a general filler noun. *Mea* was probably used as a filler form for verbs of speaking, see the comparative examples in (1) and (2). It is possible that this use occurred so frequently that in that context, it was taken to mean ‘say’, rather than ‘do’.

1. He went (was like) “Never”

2. Ka mea ia “Korekau”
   TAM say IIISG Never
   ‘He did “Never”’

I thus conclude that for modern Māori, despite the etymology, there is a case to

be made for considering the ‘say’ and ‘thing’ senses as different lexemes.

5.2 Grammatical review

This section will look at the grammatical functions of *mea* and whether the grammar might support making one lexeme distinct from the other. There are obvious grammatical distinctions between the senses ‘thing’ and ‘say’, that is, one is nominal and the other verbal in sense. The three labels given for nominal and verbal types from the grammar review are noun, indefinite pronoun and canonical transitive verb. The different types of nominal functions include *he mea* clefting, and indefinite pronouns. There are also those forms of *mea* which occur in conjunctions introducing cause clauses, manner clauses, condition clauses, and clauses of time. Apart from stem noms, the ‘thing’ sense behaves like a noun, and is preceded by determiners, while the ‘say’ sense behaves like a verb, and is preceded by TAMs. This is equivalent to the two having different word-forms, and under Lyons’s third criterion is grounds for saying that they are different lexemes.

Bauer (1997:589-608) discusses various types of adverb clauses. We will begin by looking at cause clauses as they were quite frequent in the MBC. The conjunctions introducing cause clauses listed in the dictionary review, all meaning ‘because’, are:

3.	i	te	mea	(ai)
	cause	theSG	thing	part.
4.	nā	te	mea	
	belong	theSG	thing	
5.	nō	te	mea	
	belong	theSG	thing	
6.	tā	te	mea	
	belong	theSG	thing	

The conjunctions which are listed by Bauer (1997:600) but not by the dictionaries are: *i te mea hoki* and *he mea hoki*. Bauer (1997:600-601) states that all instances of *mea* in these conjunctions are the ‘thing’ sense and all function as subordinating conjunctions; another important note that Bauer makes is that the preposition is often omitted in casual speech, which is a

regular occurrence in the MBC.

In manner clauses, Bauer (1997:598) states that the most common conjunction is *me te mea* which is sometimes followed by *tonu* or *nei*. There were not any instances of *me te* co-occuring with the 'say' sense of *mea* in the MBC. Therefore, it is possible this expression can be tagged automatically for the 'thing' sense. In an attempt to test this hypothesis, I asked a native speaker whether it is possible to say the sentence in example (7). I then removed the pre-verbal modifier ***āta*** and asked if it still made sense. My informant felt that perhaps *me tana mea mai* lit 'and his say hither' would be more appropriate in this context.

7.	Ka	ara	ia	i	tana	moenga	me	te
	TAM	arise	IIISG	DO	his	bed	and	theSG
	āta	mea,	"kāore	ahau	i	tino	mate"	
	carefully	say,	NEG	ISG	TAM	very	ill	

'He rose from his bed and meekly said "I'm not that sick".

Though there were not any occurrences in the MBC, condition clauses were listed as a possible grammatical environment in the dictionary review. Bauer (1997:603) states that these types of clauses are most frequently introduced by *mehemea.* Due to their absence from the MBC these are not discussed further.

The next issue that arose was regarding the form *he mea* and the possible ambiguity one might find in this context. It is possible for the form to either be used in *he mea* clefting in which the sense of *mea* is 'thing', or used in a stem nominalisation with the verbal sense *mea* preceded by the determiner *he*. The environments of both forms are looked at here in order to differentiate the sense of *mea* in each construction.

He mea clefting was also found quite regularly in the MBC as illustrated in (8) and (9). Bauer (197:536) notes that cleft sentences with *he mea* exhibit charactaristics of ergative marking. Example (8) shows the verb as active in form and the agent marked by *e*. The subject NP is unmarked as is the usual marking of subject NP's. The characteristics contained in this type of construction are significant in identifying the sense 'thing' from the sense 'say'.

8. He mea pupuri tana ringaringa e taua minita
   CLS thing hold his hand by that minister
   'His hand was held by that minister' (more lit. 'A thing that was held by the minister was his hand')

Example (9) is another sentence taken from the MBC with the same ergative structure as (8):

9. He mea waru hoki e au
   CLS thing peel also by 1SG
   '[they] were peeled by me'

There are examples in the MBC which are marked differently but still appear to be *he mea* clefts. Examples (10) and (11) show a different type of construction which excludes the patient from the phrase of the canonical transitive verb *kimi* 'search' in (11). Examples (10) and (11) demonstrate the use of *n-* possessive marking of the agent.

10. He mea kōrero anō nā Rewa
    CLS thing say again N-FORM Rewa
    'It was mentioned again by Rewa'

11. Homai taku taonga he mea kimi nāku
    give my1SG treasure CLS thing search N-FORM1SG
    i te whakarunga
    PREP the3SG above
    'Give me my treasure that was searched for by me up above'

However, not all sentences beginning with *he mea* are clefts; some of them are stem nominalisations of the 'say' sense. Example (12) shows a stem nominalisation co-occurring with the manner particle *noa* 'freely' which is far more likely to occur with a verb than a noun (Bauer, 1997:338)

12. He mea noa ake nōku kia tukuna
    CLS say freely up N-FORM 1SG TAM release
    ngā moni e te kaitiaki
    thePL money by the3SG caregiver
    'It was freely stated by me to have the money released by the caregiver'

Stem nominalisations (discussed in Terminology 2.6) can be cause for

ambiguity of the sense of a phrase in two ways. Firstly the determiner *te* can co-occur with *mea* to form stem nominalisations and the determiner *he* can co-occur with *mea* to form a stem nominalisation. The form of stem nominalisations mirrors that of nominal senses of *mea* preceded by these determiners. The way in which we can differentiate between the various forms and their senses will be looked at shortly.

5.3 Results of *mea* from analysis of MBC

There are a total of 6,062 occurrences of *mea* in the MBC which accounts for 6.0% of all tokens. It is the 27th most frequent item in the MBC.

After the analysis of *mea* was complete, three lexemes attached to the word mea surfaced from the raw data in the MBC. Those three lexemes were: 'thing', 'say' and 'mayor'. The lexeme 'thing' had the highest frequency followed by 'say' then 'mayor'. The following table shows the frequency of these items.

Table 5.4. Raw frequency results for senses of *mea*

	'thing'	**'say'**	**'mayor'**	**Total**
No. of Tokens	5,101	695	10	5806

The multi-word units of *mea* (such as the conjunctions discussed in 5.2) were analysed as a unit and therefore are not represented in Table 5.4 above. There were a total of 1007 multi-word units of *mea*.

The following section outlines the process of analysis of the lexemes *mea* 'thing', *mea* 'say' and *mea* 'mayor'. It will also look at the different environments in which *mea* occurs. An attempt is made to establish a set of syntactic guidelines to identify the two main meanings 'thing' and 'say' in a corpus, thus enabling a tagger to effectively tag for each lexeme.

Firstly the grammatical features of the lexemes *mea* 'thing' and *mea* 'say' will be examined in order to establish those environments that these words systematically appear in based on the grammatical function of each lexeme. Secondly, the functional categories will be analysed such as: the phrase-type

markers that each lexeme co-occurs with and their relative distribution; the modifiers that occur in the phrase and their co-occurrence with each lexeme; and the sentence and clause position of each lexeme will also be looked at in order to conclude what the position might tell us about the sense of the word. Those items which are not easily distinguished by the previous two steps will be further analysed by the wider context and also by any patterns in the collocates which occur. The results of these processes will then be discussed and final conclusions will be made about how successful these processes were in distinguishing each lexeme.

5.3.1 Functional Categories

In this first stage of analysis, other syntactical elements have been looked at for each example of *mea* that is, sentence/clause position and other contextual clues such as any modifier in the phrase. Where *mea* functions as the lexical head of a *he mea* clefting construction, the patient and agent phrase (when present) has also been noted.

5.3.2 Phrase type markers – TAMs

The first step is to consider syntactic criteria such as co-occurring phrase type markers which indicate a verbal use of *mea* or a nominal use. So for example where the phrase type marker *ka* occurs preceding *mea* we can automatically assign that to a verbal sense, eliminating the nominal ‘thing’ sense.

Table 5.5 shows a breakdown of the TAMs which co-occurred with the lexeme *mea* ‘say’ from most frequent to least. The TAMs that preceded *mea* ‘thing/do’ have also been given for comparison.

Table 5.5. Tense Aspect Mood Markers

TAM	**'thing/do'**	**'say'**	X2	2-tailed P	sig
ka	5	222	1582.376	>0.0001	extremely
e … ana		168	1232.000	>0.0001	extremely
e		73	535.333	>0.0001	extremely
i	3	44	296.480	>0.0001	extremely
kei te		17	124.667	>0.0001	extremely
Ø		80	586.667	>0.0001	extremely
kua	1	48	342.8606	>0.0001	extremely
kia	1	6	36.019	>0.0001	extremely
i te		1	7.333	0.0068	very
kei		1	7.333	0.0068	very
me	2	15	93.561	>0.0001	extremely
Total:	**12**	**675**			

The rows with the result 'extremely' significant correspond with two-tailed P on

the basis of my null hypothesis about ‘expected values’. It is clear that the occurrence with any TAM strongly predicts the ‘say’ sense. This is scarcely surprising, since it is verbal, and the ‘thing’ sense is nominal.

However, the nature of the spoken corpus slightly reduces the usefulness of this type of analysis, because *mea* ‘thing’ is occasionally used as a substitute for the verb in a verb constituent marked by *ka* and various other TAMs. This was not a regular occurrence (there were only 12 cases in the MBC) and should be even less frequent in data from non-spoken sources. *Mea* ‘thing’ co-occurs with the TAMs *ka*, *i*, *kua*, *kia* and *me*. It appears that where there is a pause in thought the speaker uses *mea* to fill the verbal slot whilst searching for the correct verb. Examples (13-15) show various examples which are typical of this use of *mea*.

13. hei huarahi mō ngā tauira,
PREP pathway for thePL student
kāore i mea, i taea ki te
NEG TAM thing TAM able to theSG
taumata tino teitei
level very tall
‘as a path for the students who aren’t whats-it, able to [excel] to their fullest potential…’

14. Ā, he wā ka mea ka huri
Ah, CLS time TAM thing TAM turn
mai wērā mahi a, o tērā tangata
hither that work ah of that person
‘And one time, which will something, which will change that job, ah, of that person’

15. kua mea, kua tae noa mātou
TAM thing TAM arrive freely IIIPLEXCL
‘…has something’d, we have returned’

Those examples which had Ø TAM were identifiable as the ‘say’ sense due to the directional particles *mai* and *atu* following *mea*. Where there was *ana* following *mea* this also indicated the ‘say’ sense.

All other examples exemplified verbal uses of *mea* ‘say’. So 1.7% of those

TAMs that co-occurred with *mea* were the ‘thing’ meaning and 98% were the ‘say’ meaning. I think in a corpus of non-spoken data, we would find that we could be 100% certain that where *mea* co-occurs with a TAM, it is the ‘say’ sense.

5.3.3 Determiners

The next step was to look at what the determiners as the phrase-type marker might suggest regarding the meanings *mea* ‘thing’ and *mea* ‘say’. Table 5.6 shows the frequency of the determiners that co-occurred with the senses ‘thing’, ‘say’ and ‘mayor’.

Table 5.6. Determiners with *mea*

Det	**'thing'**	**'say'**	**'mayor'**
te	950	12	10
he	317	2	0
taku	8	2	0
tana	2	4	0
hei	2	1	0
tō	0	1	0
other	2800	0	0
Total	**4079**	**23**	**10**

The two-tailed P calculations do not shed any further light on the results and so have not been included in Table 5.6. The results from Table 5.6 are as expected and show a high frequency of determiners as strong identifiers for the 'thing' sense.

The ten occasions where *Mea* 'Mayor' occurred in the corpus were all capitalised, so therefore a determiner preceding *Mea* with a capital letter could be tagged as 'mayor' (but the reverse is, of course, not true: we cannot

conclude that det + *mea* without a capital letter is not the ‘mayor’ sense; it could be). ‘Mayor’ could also be tagged for the environment of *e* ‘vocative’ preceding *Mea*.

The lexeme *mea* ‘say’ co-occurred with the determiner *te* 12 times in the MBC. Of these 12 examples, 6 were in *ki te* complements. These types of constructions are labelled as verbal *te* + stem forms by Bauer (1997:529). In these types of constructions the subject is deleted under identity with the subject in the matrix clause. The *ki te* complement in example (16) is functioning as a purpose adverbial which is additional information added to the matrix clause *ka haere atu au ki kō*. The *ki te* complement has the goal phrase *ki a ia* which could not occur following nominal forms preceded by *ki te*. Example (16) also demonstrates an example of the canonical transitive verb *mea* with a direct object which implies verbal output. Both the direct object and goal phrase are useful indicators of the ‘say’ sense.

16.	ka	haere	atu	au	ki	kō,	
	TAM	go	away	ISG	to	over there	
	ki	te	mea	atu	ki	a	ia,
	to	theSG	say	away	to	PERS	IIISG
	arā,	āwhea	te	parakuihi?			
	that is	when	theSG	breakfast			

‘I went over there to say to him, that is, when is breakfast?’

There were two instances of stem nominalisations not preceded by *ki* in examples (17) and (18). The directional particle *atu* occurred following *mea* in example (17) and the directional particle *mai* in (18). Example (17) exemplifies the stem nominalisation as a complement of the neuter verb *mau*. The stem nominalisation in (18) occurs in the predicate head of an equative sentence which has a *ko*-fronted subject.

17.	…te	mea	atu	koe,	ō,
	…theSG	say	away	IISG	um
	e	whakapai	ana	au	
	TAM	clean	TAM	ISG	

‘you said, um, that I was cleaning up

18. ko te mahi ko te mea mai
PREP the work PREP theSG say hither
'work said …'

There were two examples of the lexeme *mea* 'say' that co-occurred with the determiner *te* in a *taea* complement. These examples can be differentiated from the 'thing' sense by searching for *taea* as a left collocate. There was one example that did not have a directional particle functioning as a post-posed modifier in (19):

19. e kore e taea e au te
PART nil TAM able by ISG theSG
mea he hapu anō a Te Waiariki
say CLS sub-tribe also PERS Te Waiariki
me Ngāti Kororā nō Ngāti Wai…
PREP Ngāti Kororā from Ngāti Wai
'I wasn't able to say that Te Waiariki and Ngāti Kororā from Ngāti Wai were also [my] hapu'

Again, the nature of the corpus affected the data output; there were occasions throughout the entire sample set of *mea* where prepositions had been deleted as in example (20). There were only two instances where the determiner *te* preceded *mea*. Bauer (1997:600-601) states that in cause clauses the preposition is sometimes elided, and it appears that it is omitted in other phrases also.

20. te kōrero mai ki a au,
theSG talk hither to PERS ISG
te mea mai ki a au, āe,
theSG say hither to PERS ISG yes
'[he/she] said to me, yes'

Another instance which arose due to the spoken nature of the data was the formal address from one to another as in example (21):

21. …e te mea mai, e Pita haramai
…voc theSG say hither voc Peter come hither
'[it] is said, come, Peter'

There was only one instance of *mea* co-occuring with the determiner *tō* which is a t-possessive and it was followed by the directional particle *atu*.

All cases of the 2,800 determiners that remained were followed by *mea* ‘thing’. There was no co-occurrence of either *mai* or *atu* in the phrase with the meaning ‘thing’. There were, however, examples of the ‘thing’ meaning with the directional particle *ake* ‘upwards’ which, however, never co-occurred with a TAM.

It thus appears that there are usually other features present when the lexeme *mea* ‘say’ occurs with a determiner which can be used to identify the appropriate sense, and that, because these cases are relatively infrequent, the presence of a determiner in the phrase periphery is strongly indicative of the lexeme *mea* ‘thing’.

5.3.4 Modifiers

In all cases the directional partcles *mai* ‘hither’ and *atu* ‘away’ functioned as post-posed modifiers to *mea* ‘say’. This was significant in the identification of the lexeme *mea* ‘say’ as there were no cases where the meaning ‘thing’ was modified by these directional particles.

All prepositional phrases that functioned as modifiers to *mea* were the ‘thing’ lexeme. All cases of ordinal numbers following *mea* were the ‘thing’ lexeme. There was a high frequency of *nui* following *mea*, and all cases were the ‘thing’ lexeme.

5.4 Conclusions

The information gathered in the dictionary review and etymology review did not support the senses ‘thing’ and ‘say’ as being distinct lexemes. However, using Lyons’ first criterion regarding native speakers’ intuition, 10 informants all agreed that the sense ‘say’ and the sense ‘thing’ were indeed separate lexemes. This was supported by the fact that the two occur in largely different syntactic environments.

The results of the analysis of the raw data provided the following set of rules that could be applied to tagging a corpus for the lexemes *mea* ‘thing’ and *mea* ‘say’:

1. If mea occurred in a multi-word conjunction = mea ‘say’
2. If TAM precedes mea = ‘say’ sense
3. Ø marking preceding mea = ‘say’ sense
4. If DET precedes mea = ‘thing’ sense unless followed by mai or atu
5. If ki te precedes mea = ‘thing’ sense unless followed by mai or atu
6. If cause conjunction = ‘thing’ sense
7. If he mea = ‘thing’ sense when followed by a verb

6 Tau

6.0 Introduction

The chapter begins with a survey of the various types of information about *tau* which need to be considered in determining how many lexemes are realised by this word-form. When the lexemes have been determined, an analysis of the environments in which they occur will be given in the results section. This leads to a conclusion about how these lexemes might be tagged in a corpus.

6.1 Establishing lexemes associated with *tau*

This section includes information about the word *tau* from dictionaries and etymology reviews in an attempt to differentiate the associated lexemes.

6.1.1 Dictionary review

The dictionary review in Table 6.1 summarises how the dictionaries analyse the lexemes realised as *tau.* The information has been placed in the following order: meaning, associated grammatical functions and senses. The meanings differentiated in the Table are those recognised in the dictionaries, so that if a dictionary lists two words under the same head entry, both words are entered here under the same meaning. Similarly the senses used here are those associated with a particular head word by the dictionary in question. The associated grammatical functions are abbreviated as follows: Noun (N), Transitive Verb (VT), Action Intransitive Verb (VI), State Intransitive Verb (VS). The dictionary column will contain a tick in the row for the head word if the dictionary has entered this as a head word. As mentioned in the dictionary review in Methodology, *HPK* and *Williams* are sometimes identical, and this is one occasion where *HPK* has not provided any additional information.

Table 6.1 Information from dictionaries about *tau*

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Kimihia Tirohia***
tau¹ *Year*	√	√	√	√
Associated grammatical functions:	*N*	*N*	*N*	*N*
Senses:	*year, season*	*year, age*	*season, year, period of time, interval*	*year*
tau² *to settle*		√	√	√
Associated grammatical functions:		*VT, VI, VS*	*VI*	*VI*
Senses:		*(-ria) to land, alight, come to rest, settle on, count, settle, perch, ride at anchor;* *(no passive given for this entry) to settle down, subside, abate;* *(stative) be neat, comely, smart, attractive, handsome, becoming, suitable, beautiful, cute, befitting*	*to alight, come to rest, fall of blows, come to anchor, lie to, ride to anchor. Float, settle down, come over, supervene (of feelings), lie steeping in water, be suitable, be comely, befit, be possible, be able*	*to alight, come to rest, fall, be suitable, befit*

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Kimihia Tirohia***
tau[3] *Lover/Spouse*		√	√	√
Associated grammatical functions:		*N*	*N*	*N*
Senses:		*husband, spouse, partner, lover, darling, beau, boyfriend, girlfriend, sweetheart*	*lover/spouse/darling*	*lover, spouse*
Tau[4] *String of garment/loop*		√	√	
Associated grammatical functions:		*N*	*N*	
Senses:		*string (of a garment), loop or thong (of a patu)*	*string of garment, loop or thong of mere.*	

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Kimihia Tirohia***
Tau[5] *Ridge of a hill*		√	√	
Associated grammatical functions:		*N*	*N*	
Senses:		*ridge*	*ridge of a hill, reef*	
Tau[6] *Sing/song*		√	√	
Associated grammatical functions:		*VT, VI, N*	*VT, VI, N*	
Senses:		*(-a) to sing, bark (of a dog). song, chant at the beginning of a speech*	*sing, sing of, bark, song, noise, report*	

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Kimihia Tirohia***
Tau[7] *Attack*		√	√	
Associated grammatical functions:		*VT*	*VT*	
Senses:		*(-ia,-ria) to attack*	*attack*	
Tau[8] *Awesome![my gloss]* *(expressing satisfaction)*			√	
Associated grammatical functions:			*VS*	
Senses:			*Yay!*	

Dictionary	***Biggs***	***Moorfield***	***Williams***	***Kimihia Tirohia***
Tau[9] *Number*		√		√
Associated grammatical functions:		*N*		*N*
Senses:		*number*		*number*

In Table 6.1 Biggs, Moorfield, Williams and *Tirohia-Kimihia* all agree that *tau* ‘year’ is a distinct lexeme. The senses that all the dictionaries have attached to this lexeme are ‘year’ and ‘season’. Moorfield includes the sense ‘age’ and *Williams* includes two other senses associated with the lexeme ‘year’, that is ‘period of time’ and ‘interval’. *Tau* ‘year’ is the only entry for *tau* in Biggs.

Moorfield, Williams and *Tirohia-Kimihia* all include the lexemes *tau* ‘settle’ and *tau* ‘lover/spouse’. In addition Moorfield and Williams list the following lexemes: *tau* ‘string of garment/loop’, *tau* ‘ridge of a hill’, *tau* ‘sing/song’, *tau* ‘attack’. Williams also lists *tau* ‘awesome’; however Moorfield lists this in a separate entry with the multi-word unit *tau kē nei* ‘awesome’. Moorfield and *Tirohia-Kimihia* (the two most modern dictionaries) both include an entry for *tau* ‘number’.

Williams includes under the sense ‘string of garment, loop or thong of mere’ the phrases *tau o te ate* and *tau o te manawa* ‘heart strings, deep emotion’. Moorfield also notes *tau o te ate* in a separate entry with the same sense. These have not been included as a sense of *tau* here as they can be searched for as a unit in the MBC and therefore considered multi-word units.

The grammatical information provided by the dictionaries will be considered in 6.1.3.

There are nine separate entries given for *tau* by the dictionaries listed in Table 6.1, and thus nine possible lexemes associated with *tau* in Maori, although only one dictionary includes all 9: ‘year’, ‘settle’, ‘lover/spouse’, ‘string of garment/loop’, ‘ridge of a hill’, ‘sing/song’, ‘attack’, ‘awesone’ and ‘number’. However, the majority of the dictionaries only record three: ‘year’, ‘settle’, and lover/spouse’. One reason for such difference between the selection of lexemes by each dictionary is that it reflects differences in the frequency of these senses. *Kimihia Tirohia* selected lexemes of high frequency. The frequency of these lexemes is analysed further in Section 6.2.

6.1.2 Etymology

Greenhill & Clark (2011) has six separate entries for *tau* and Tregear (1891) has eleven senses for *tau* listed under one entry. Table 6.2 lists the six separate entries from Greenhill & Clark (2011) in the first column and the senses from

Tregear in the second column. Because Tregear does not separate senses into lexemes, I have aligned the senses he gives with the separate entries from Greenhill & Clark (2011) for comparison.

The senses of *tau* recorded by Tregear (1891) and Greenhill & Clark (2011) in Table 6.2 are very different. Greenhill & Clark (2011) does not list ‘to bark as a dog’; ‘door’; ‘to lie at anchor or moorings’; or ‘to attack’, while Tregear does not list ‘hang suspended’; ‘tie in bunches’; ‘thread on string’ and ‘count’. The senses which are missing from these two comparative dictionaries that the dictionaries in Section 6.1 include are ‘lover/spouse’ and ‘awesome’.

Table 6.2 Senses of *tau* listed by Tregear (1891) and Greenhill & Clark (2011)

Greenhill and Clark (2011)		*Tregear (1891)*	
1.	*season/year*	*1.*	*year*
2.	*loop of cord attaching a club to the wrist; cord handle of a basket*	*2.*	*the string of a garment; a loop or thong*
3.	*reef, ridge of a hill*	*3*	*the ridge of a hill*
4.	*sing, song*	*4.*	*a song; to sing*
		5.	*to bark, as a dog*
		6.	*a door*
		7.	*the carved stern-piece of a canoe*
5.	*settle, as a bird, anchor, as a boat, come to rest*	*8.*	*to alight upon; to rest*
		9	*to lie at anchor or moorings*
.6	*be able, suitable*	*10.*	*to be suitable, to become, to look well*
		11.	*to attack*

The dictionaries reviewed in Table 6.1 did not include the following meanings from Table 6.2 as being associated with *tau*: ‘a door’ and ‘the carved stern piece of canoe’. Despite this, if we look at the word in Māori for ‘a door’, the majority of dictionaries agree that the lexeme for this sense is *tatau*, which is a reduplication of *tau*. Tregear provides an example of the lexeme ‘door’ as *tau* in Māori, however.

The sense ‘awesome’ appears to be a multi-word unit according to the dictionary review and there is no mention of this sense being attached to *tau* in the etymology review. Moorfield includes the multi-word units *tau kē nei* ‘cool/neat’; *(te) tau/kino (kē) (hoki)*; *(ka/he/te) tau/kino (kē) (hoki)*; *(ka) tau/kino (kē) (hoki)*. These are possible combinations using *tau* to create the meaning ‘awesome’ [my gloss]. The multi-word units *tau kē* and *tau kē nei* are the most commonly associated combinations with the meaning ‘awesome’ given by the dictionaries. It is also noted in Moorfield that if *te* ‘the’ is omitted *kē* must follow *tau* in these expressions.

The dictionaries and the etymology sources differ with regard to what constitutes a distinct lexeme of *tau*. The dictionaries consider all the following senses to be related to a single lexeme: ‘to alight upon’, ‘come to rest’, ‘to anchor’, ‘to be attractive’, ‘to befit’, ‘be possible, be able’. Yet the etymological sources enter these senses separately as ‘to alight upon’, ‘to rest’; ‘to lie at anchor or moorings’; ‘to be suitable’, ‘to become’, ‘to look well’.

The way in which these senses have been listed may suggest that *tau* is more complicated to deal with than it seems, because there is no clear division between polysemes and homonyms. Therefore it is not clear what constitutes an individual lexeme from these lists. What we can do, however, is look at the cognates of these words throughout Polynesia and see if this gives any indication of how we might separate the senses into lexemes.

Overall, only six of these senses occurred in the MBC: *tau* ‘year’, *tau* ‘settle’, *tau* ‘awesome’, *tau* ‘love’ *tau* ‘song’ and *tau* ‘number’. Due to the very low frequency of *tau* ‘love’, *tau* ‘song’ and *tau* ‘number’ in the MBC, these words have been excluded from the comparative Polynesian etymology review that follows.

Combining the information from Tregear and Greenhill & Clark (2011), *tau* ‘year’ is found in Samoan, Tahitian, Hawaiian, Tongan, Rarotongan, Mar quesan, Mangarevan and Pukapukan as well as in Māori. From the same sources, *tau* ‘settle’ and/or ‘anchor’ is found in Samoa, Tahitian, Tonga, Rarotonga, Marquesan, Moriori, East Futuna, East Uvea, Pukapukan, Takuu, Tikopia, Tuvalu, West Uvea. Yet in Hawaiian, the form is *kau*, in Mangarevan

the sense is ‘to land’, in Kapingamarangi and Nukuoro, the form is *dau* and the sense ‘land’, in Vaekau-Taumako the sense is ‘arrive, got to’. The sense ‘be able, suitable’ is found in Samoa, East Futuna, Easter Island, Rarotonga, Tonga, Moriori, Mangareva, Pukapukan, Takuu, Tuvualu and West Futuna. Greenhill & Clark (2011) lists the sense ‘habitual action’ for Samoa, but Tregear lists ‘right, proper, fit, to be right and proper’. Takuu has the meaning ‘equal to a task; able to do something, enough, sufficient; (of clothes) fit’. West Futuna has ‘follow in the ways of, take after, learn from’. Cognates for ‘loop of rope’ in Māori are found in Samoa, East Futuna, Easter Island, East Uvea, Luangiua, Kapingamarangi, Tahiti, Hawaiian, Rarotongan, Tongan, Moriori, Marquesas, Mangareva and Pukupukan. The form in Hawaiian and Luangiua is *kau* and in Kapingamarangi is *dau*. Easter Island, Tahitian, Marquesas, Pukapukan and East Uvea have the form *tautau*. The majority of these languages have the sense ‘to hang, to hang upon’. The lexeme ‘to hang, to hang upon’ is listed in Māori in Williams and Moorfield as *tautau* yet the lexeme *tau* ‘loop of rope’ is listed as a separate lexeme.

The evidence from cognates in other Polynesian languages gives support to recognising the following clusters of senses as belonging to separate lexemes: ‘year’, ‘settle, anchor, land’, ‘be able, suitable’, and ‘loop of rope’.

6.1.3 Grammatical review

The following section will look at the grammatical functions of *tau* and how the grammar might assist in separating lexemes.

The dictionary review and etymological information was not as decisive in distinguishing lexemes as it was for *kī*. It is now that we look to the grammatical information to see what we might glean from this information in terms of separating lexemes. It is here that Lyons’s third criterion for absolute homonymy, namely, ‘grammatical function’ will be applied to analyse lexemes. Lyons (1977:22) states that under the third criterion of distinction of lexemes, formal identity and grammatical equivalence must not be present.

The grammatical functions given by these various sources are as follows. The sense ‘year’ is classed as a noun by Biggs, Moorfield, Williams and *Kimihia Tirohia.* The sense ‘lover/spouse’ is labelled as a noun, as is ‘ridge of a hill’,

‘string of garment/loop’, ‘song’ and ‘number’. The sense ‘settle’ is classed as an action intransitive verb by Moorfield, Williams and *Kimihia Tirohia*, though Moorfield also includes a passive ending *(-ria)* for the ‘land, to light, to come to rest’ sense which then qualifies *tau* to be classed as a canonical transitive verb. The label stative, which we refer to as a state intransitive, has been assigned to the sense ‘be neat, comely, smart’ by Moorfield. The sense ‘sing’ has been labelled a canonical transitive verb by Williams and Moorfield as has the sense ‘to attack’. The sense ‘bark’ has been labelled as an action intransitive verb by both these dictionaries.

Let us examine the lexemes suggested by the dictionary and etymology reviews. If we consider *tau* ‘year’, *tau* ‘sing/song’, *tau* ‘settle’, *tau* ‘be able/suitable’, *tau* ‘attack’ and *tau* ‘bark’ we could claim that equivalence in grammatical function does not exist. This would suggest that we treat them as separate lexemes.

The grammatical function assigned to *tau* ‘year’, is noun, while *tau* ‘settle’, *tau* ‘be able/suitable’, *tau* ‘attack’ and *tau* ‘bark’ are verbs. *Tau* ‘sing/song’ falls under both categories with ‘sing’ a verb and ‘song’ a noun. If we analyse the grammatical functions of these verbs further there are more grammatical distinctions to be made as to their verb types. *Tau* ‘settle’ is classed an action intransitive verb as is ‘bark’; *tau* ‘sing’ is considered a canonical transitive verb as is ‘attack’; *tau* ‘be able, suitable’ is labelled a state intransitive verb. There is difference in opinion between Williams and Moorfield regarding the grammatical function of *tau* ‘to land’. Williams lists this sense under the same head entry as the sense ‘settle’ as does Moorfield. However Moorfield considers the grammatical function to be a canonical transitive verb.

The dictionary review identified a pattern among passive suffixes assigned to the various canonical transitive lexemes. Williams and Moorfield note the passive suffix *–a* for the lexeme *tau* ‘sing’ yet have listed the passive suffixes *–ia* and *–ria* for the lexeme *tau* ‘attack’. Moorfield assigns the passive suffix *–ria* to the sense ‘to land’. Here we can claim that formal identity under Lyons’ third criterion does not exist and therefore could consider these as distinct lexemes due to the differences in their passive forms.

6.1.4 Conclusions about lexemes associated with tau

The dictionary review provided an insight as to how native speakers' intuitions discriminate lexemes. In general the dictionaries agree, though many give no information about rarer ones. Nevertheless even when senses were included in dictionaries, there were inconsistencies between whether that word was entered as a head word or whether it was listed as a sense under a different head word e.g. 'count' being listed under the head word 'settle' by Moorfield. The dictionary review did not offer clear-cut divisions between lexemes.

The etymology review provided slightly more insight, in that, in Greenhill & Clark (2011) there were clear divisions between lexemes. However those lexemes that posed problems in the dictionary review, i.e. 'lover/spouse' and 'awesome' were not listed, and so no help is available from this source. One very important pattern that emerged from both the dictionary and etymology reviews was that the lexemes in Greenhill & Clark (2011) were all grammatically distinct. The grammatical functions given by the dictionaries align for the most part with the grammatical divisions between lexemes in Greenhill & Clark (2011). In Table 6.2, we see that the six lexemes listed 'season/year'; 'loop of cord attaching a club to the wrist; cord handle of a basket'; 'reef, ridge of a hill', 'sing, song' and 'settle, as a bird, anchor, as a boat, come to rest' and 'be able, suitable' align with the grammatical distinctions discussed in Section 6.1.3. Lyons' third criterion thus provides strong evidence as to what we might consider as distinct lexemes amongst the verbs. When we turn to the nouns, we can see that the senses are so semantically diverse that there is no likelihood of them being related. We then find that we have eight distinct lexemes: *tau* 'year', *tau* 'loop of cord', *tau* 'reef, ridge of a hill', *tau* 'settle', *tau*, 'sing/song', *tau* 'to bark', *tau* 'be able/suitable' and *tau* 'to attack'.

The sense 'awesome' has not been considered a distinct lexeme due to its function in a multi-word unit which is conditional upon other particles for this meaning. Though Williams lists the meaning of *tau* as a distinct lexeme with the sense 'awesome' it is exemplified as *tau!* which could be tagged with an exclamation mark. The etymology review did not acknowledge the sense 'awesome' at all. Moorfield lists the multi-word unit and recognises *tau kē nei* as 'cool, neat' therefore supporting this meaning of the word in this context as a

distinct lexeme. The term ‘lover/spouse’ was not mentioned in the etymology review, yet this lexeme surfaced from the dictionary information. There were two examples of ‘love’ in the MBC in (1) and (2), where example (1) was used as an address term, and is probably more likely an English-influenced translation like the address term ‘love/sweety’ as opposed to ‘lover/spouse’. The second example was in formal speech to acknowledge those who have passed on, so was not in casual usage. It also did not contain the ‘lover/spouse’ sense, but more the sense of ‘precious one’. Example (2) shows the use of *tau kahurangi* – there is a similar meaning in Moorfield for *tau kahurangi* which is translated as ‘honourable lover’. These were the only two instances of ‘love’ in the MBC.

1.	Ka	kī	mai	ngā	wāhine	ki	a	au,
	TAM	say	hither	thePL	woman	to	PERS	ISG
	akona	mai		mātou	ki	te		karanga.
	teach PASS	hither		IPLEXCL	to	theSG		call
	E	tau,	kāore	au	e	mōhio		
	VOC	love	NEG	ISG	TAM	know		

‘The women say to me, teach us to call. Oh love, I don’t know how…’

2.	Ko	taku	tau	kahurangi	tērā
	PREP	mySG	love	precious	that

‘That is my precious love’

6.2 Results of *tau* from the analysis of the MBC

There are a total of 3,096 occurrences of *tau* in the MBC and it accounts for 3.0% of all tokens. It is the 56th most frequent item in the MBC.

After the analysis of *tau* was complete, 7 meanings of *tau* were found in the MBC: *tau* ‘year, *tau* ‘settle’, *tau* ‘be fitting, suitable’, *tau* ‘awesome’, *tau* ‘number’, *tau* ‘love’, and *tau* ‘song’. The following table shows the frequency of these items. This excludes the 92 instances of the proper noun *Tau.* The total number of tokens represented in Table 6.3 is less than the total number in the MBC due to the exclusion of proper nouns and unusable examples from the data.

Table 6.3 Raw frequency results for senses of *tau*

	'year'	***'settle'***	***'awesome'***	***'number'***	***'love'***	***'song'***	***Be fitting***	***Total***
No. of Tokens	*2404*	*464*	*2*	*1*	*2*	*3*	*5*	*2881*

The following section outlines the process of analysis of the senses *tau* 'year', *tau* 'settle' and comments on the five other senses.

6.2.1 Structures

The grammatical distinction between lexemes provides a good framework to begin the analysis of the environments of these lexemes. The first clear case is the division into nouns and verbs. We can begin by looking at the phrase-type markers which will automatically provide us with those that are preceded by determiners, and those that are preceded by TAMs. Those lexemes that are preceded by determiners are highly likely to be nouns and those lexemes preceded by TAMs are definitely verbs. The cases where ambiguity may arise are those where verbs occur as stem nominalisations.

We will begin by looking at the grammatical environments of the lexemes established in Section 6.1.3 of this study. Firstly, let us consider the nominal environments of 'year', 'ridge of a hill', 'song' and 'loop of cord'. These lexemes will be found as the lexical head of nominal predicates, prepositional phrases and subject noun phrases. They are not related semantically, and this means that it is likely that their context will distinguish them in a corpus. It would be expected that there would be obvious contextual differences that signal the less frequent items, such as example (3), where *o ō kahu* 'of your clothes' is an obvious phrasal collocate which might signal the lexeme 'string of garment'. In example (4) *o te patu* 'of the weapon' is another good indicator of the 'loop of cord' sense. Example (5) contains the collocate *maunga* 'mountain' which precedes the sense 'ridge of hill' and the action intransitive verb and adverbial expressing goal including the prepositional phrase in which 'ridge of hill' occurs, that is *ka haere i runga,* is a clear signal for the 'ridge of hill' sense: you wouldn't ascend a song or loop of cord and so on. *'Song'* has an obvious collocate

‘*waiata’* which would in most cases be found in close proximity in various grammatical functions.

3.	Wetea	te	tau	o	ō	kahu
	unravel PASS	theSG	string of garment	of	yourPL	clothes
	‘Unravel the rope of your clothes’					

4.	Whakawiria	iho	te	tau	o	te
	twist PASS	down	theSG	loop of cord	of	theSG
	patu	ki	te	ringa		
	weapon	with	theSG	hand		
	‘Twist downward the loop of cord of the weapon with the hand’					

5.	Ka	tae	ki	runga	ki	te	maunga	nā
	TAM	arrive	to	top	to	theSG	mountain	now/then
	ka	haere	i	runga	i	te	tau	
	TAM	go	prep	top	on	theSG	ridge of hill	
	‘Arrive on the mountain, now go by the ridge of the hill’							

Another key environment for *tau* ‘year’ was the occurrence of numerals as in (6), and the question word *hia* ‘how many’ and was found as a collocate in many cases. Example (7) shows *tau* in a numeral phrase in the fronted time adverbial. There were a high number of fronted time adverbials that contained *tau* year and this was key to signalling its environment. Example (7) shows *tau* directly following the numeral. This was also a regualar occurrence in the MBC. Numeral analysis in Māori is complicated, and there are various analyses of this construction which it is not relevant to explain here (see for instance, Bauer, 1997:27)

6.	E	rua	kē	ngā	tau
	PART	two	instead	thePL	year
	‘[it] was actually two years’				

7.	E	rua	tau	au	i	reira
	PART	two	year	ISG	PREP	there
	‘I was there for two years’					

Let us now turn to the verbal lexemes and their environments. The first environment is that of the action intransitive verbs *tau* ’settle’, and *tau* ‘bark’.

Due to their grammatical equivalence, the very first and most obvious distinction to be made between the two senses, is the collocate which would be most likely to co-occur with ‘bark’, and that is ‘dog’. If the subject noun phrase contained ‘dog’, this would be a clear indicator that we have the lexeme ‘to bark’. In cases where it may not be as obvious, it is the adverbial expressing goal that is key to the ‘settle’ sense, not only in distinguishing between these two senses, but also between all other verbal lexemes. Example (8) exemplifies *tau* ‘settle’ co-occuring with an adverbial expressing goal. The adverbial *ki Aotearoa* ‘in New Zealand’ expresses the goal of the lexeme *tau* ‘settle’. Example (9) shows the locative noun *roto* functioning in the adverbial expressing goal. It was very common to find locative nouns in adverbials expressing goal with the lexeme *tau* ‘settle’. The differences between the subjects and the presence or absence of a goal phrase would be crucial in differentiating between these two lexemes.

8.	ka	tau	mai	ana	ki	Aotearoa
	TAM	settle	hither	TAM	to	New Zealand
	‘[they] will settle here in New Zealand’					

9.	ka	tau	mai	ia	ki	roto	o	Tūhoe
	TAM	settle	hither	IIISG	to	inside	of	Tūhoe
	‘he will settle within Tūhoe’							

The next grammatical environment to explore is that of the canonical transitive lexemes *tau* ‘sing’ and *tau* ‘to attack’. Canonical transitive verbs usually co-occur with direct object phrases. These direct object phrases are marked with the preposition *i.* In contrast to adverbials expressing goal marked by *ki* that sometimes co-occur with *tau* ‘settle’, the DO of the canonical transitive verbs ‘sing’ and ‘attack’ is marked by *i* and could be used to distinguish between action intransitve lexemes and canonical transitive lexemes. This is a useful way to distinguish ‘settle’ from ‘sing’ and ‘attack’. Example (10) shows the DO phrase *i te waiata* co-occuring with *tau* ‘sing’ and example (11) shows *tau* ‘attack’ functioning with the same preposition in the DO phrase. These are key to signalling the presence of one of the canonical transitive senses. Making the distinction between the two lexemes is again a matter of looking at the collocates. *Tau* ‘sing’ is most likely to occur with *waiata* ‘song’ in the DO phrase and *tau* ‘to attack’ will have words like *tāua* ‘war party’ *pā* ‘fortress’ etc.

10.	I	tau	te	koroua	i	te	waiata
	TAM	settle	theSG	old man	DO	theSG	song

‘The elderly man sang the song.’

11.	I	tau	te	tauā	i	te	pā
	TAM	attack	theSG	war party	DO	theSG	fortress

‘The war party attacked the fortress.’

The next grammatical function, which is associated with the sense of *tau* ‘be able, suitable’ is the state intransitive. In some cases, state intransitives will have an adverbial expressing cause following the predicate and or subject noun phrase. The adverbial expressing cause in example (12) is *i ngā kākahu pai*. The form of the cause phrase adverbials is similar to a DO phrase. These can be distinguished by the collocates and or context. The nature of subjects in state intransitive sentences is that of the patient and not actor. The type of subject noun phrase would also be a clear indicator for this sense. So in example (12) *tōna āhua* is clearly an inanimate thing which could not play the role of actor. Where the subject NP is an animate thing, the adverbial expressing cause again could signal this sense.

12.	Kua	tau	tōna	āhua	i	ngā	kākahu
	TAM	suitable	hisSG	appearance	cause	thePL	clothes
	pai						
	good						

‘his appearance was suitable due to his decent clothing’

Stem nominalisations are likely to cause ambiguity among these environments. When the verbal sense of *tau* ‘settle’ occurs as a stem nominalisation, the sentence may be ambiguous. Most of the time these stem nominalisations can be identified by the presence of a post-posed particle in the phrase, so for example the stem nominalisation in example (13) can be identified due to the post-posed directional particle *mai*:

13.	e	tika	ana	tā rātou	tau	
	TAM	correct	TAM	theirPL	settle	
	mai	ki	konei	i	tēnei	rā
	hither	to	here	PREP	this	away

‘it is right for them to come and settle here’

Next, a tagger needs to consider environments where *tau* functions in a multi-word unit. Moorfield (2005) lists the following multi-word units (all meaning ‘awesome’: *tau kē nei, te tau kē hoki, te tau kē nei, ka tau kē, ka tau hoki, he tau kē*. Williams, Moorfield and *He Pātaka Kupu* list the multi-word units *tau o te ate* and *tau o te manawa* ‘deep emotion’. There were not any examples like this in the corpus; however it is noted here as a possible unit to tag for this particular sense.

Another type of multi-word unit that could be tagged for *tau* ‘year’ is listed in Moorfield as *e hia N kē (mai) (nei)* ‘heaps of N’, ‘goodness knows how many N’. This construction was used quite often in the MBC as in example (14).

14.	e	hia	tau	kē	i	muri	mai,
	PART	how many	year	instead	PREP	behind	hither
	‘it was untold years afterward’						

There were some instances of this idiom that did not include the modifier *kē*. Example (15) shows an alternative form from the MBC:

15.	e	hia	tau	nāianei	kei te
	PART	how many	year	now	TAM
	haere	tonu	tāua		
	go	still	IDLINCL		
	‘What a lot of years we’ve been going for now’				

Another environment that can be tagged for the sense ‘year’ is discussed in Bauer (1997:310), that is in modifiers with linking ā- in Māori; examples (16) and (17) are from Bauer (1997:310):

16.	hui	-ā-	tau
	meeting	PART	year
	‘annual meeting’		
17.	utu	-ā-	tau
	payment	PART	year
	‘annual payment’		

Examples like (16) and (17) were not transcribed in the MBC with the hyphens in place. The reason Boyce (2006:45-46) gives for this was due to the variations of placement of the hyphens in any given text. Some texts placed the hyphen

preceding and following *ā*, other texts only following *ā* and sometimes there were no hyphens present at all. Therefore, in order to satisfy the varying placement issues, it was decided not to place hyphens in these examples at all, but to later analyse the strings in which the *ā* occurs. Boyce states that this may not have been the most productive way of transcribing as it reduced the lexical items in the data. If hyphens were present it would reduce the need to sort these types of examples in a text. *Te Taura Whiri* 'The Māori Language Commission' (Te Taura Whiri 2010:11) have orthography guidelines which specify that in these cases, the hyphen should only be placed following the *ā*.

6.2.2 Phrase type markers

As with *mea* in the previous chapter, an obvious step is to look at syntactic criteria such as co-occurring phrase type markers which indicate a verbal use of *tau* or a nominal use.

The following table shows a breakdown of the determiners which co-occurred with the various lexemes *tau* from most frequent to least. There were examples of proper names in the corpus with the determiner *a* preceding them. These were all excluded from the corpus count as a proper name can clearly be considered a different lexeme, as its form is always *Tau*.

Table 6.4 **Determiners with *tau***

Det	**‘year’**	**‘love’**	**‘settle’**	**‘be fitting’**	**‘awesome’**	**‘number’**	**‘song’**
aku	13						
taku		1					
tana/ ana	7						
ēnā	1						
ēnei	9						
ērā	5						
ētahi	10						
he	28			5			
te	874		4		1	1	1
ngā	375						
(ng)ōku	13						

ōna	24						
tā tātou	1						
tā rātou	3	2	2				
ō rātou	5						
taua	23						
tēnei/ teneki	243						
tērā, (w)ērā	126						
tētahi	6						
tō	2						
tōna	1						
Total:	**1769**	**3**	**6**	**5**	**1**	**1**	**1**

Given the tiny numbers of tokens of other senses than ‘year’, two-tailed P value statistics were unlikely to prove helpful and so were not included in Table 6.4. The results from Table 6.4 are as is expected. The majority of determiners precede the nominal sense ‘year’. Possessive determiners however are indicative of senses other than ‘year’. There were only 4 examples of the sense ‘settle’ that occurred preceded by a determiner. The first way in which the senses ‘year’ and ‘settle’ can be differentiated is to look at any post-posed modifying particle in the phrase. Most examples with the sense ‘settle’ could be identified as this sense because the directional particles *mai* and *atu* occurred in the phrase, as in examples (18-19):

18.	i	te	pūtake	o	tā rātou	tau	mai
	PREP	theSG	reason	of	theirSG	settle	hither
	‘the rationale for their arrival…’						
19.	e	tika	ana	tā rātou	tau	mai	
	TAM	right	TAM	theirSG	settle	hither	
	‘it was appropriate that they arrived here’						

Those examples that did not have directional particles had other clues signalling the correct sense, such as context words as in *o te waka whakaparaha* ‘of the broad boat’ in (20), which signals the ‘settle’ sense. Another clue diminishing the possibility of a nominal use of *tau* is the adverbial expressing goal *ki uta*. The determiner *he* is a significant signal for state intransitives in stem nominalisations as in (21). Bauer (1997:38) asserts that state intransitives occur either in verbal sentences or non-verbal sentences; the likely determining factor is whether the attribute in question is an inherent property or not. Inherent properties are expressed in non-verbal sentences. Another clue for the verbal sense is the subject noun phrase *aku karangatanga* which is not likely to occur with any of the nominal senses. The subject noun phrase in (22) *ērā āhuatanga katoa* ‘all of those aspects’ is a likely collocate for the sense ‘be fitting’ and not a likely subject noun phrase for ‘year’.

20.	te	tau	o	te	waka	whakaparaha	ki	uta
	theSG	settle	of	theSG	boat	flat	to	shore
	‘the settling of the flat boat to shore’							

21. He tau āku karangatanga o Ngāi Tahu
CLS settle myPL calling of Ngāi Tahu
'my duties to Ngai Tahu have been settled'

22. He tau ērā āhuatanga katoa
CLS settle those aspect all
'All those issues have been settled.'

Another indicator for the sense 'year' was the occurrence of *ia* 'each/every' preceding *tau* in the MBC, since all cases in this environment were the sense 'year'. However, since the other nominal senses of *tau* were so infrequent, it is not clear how strong this generalisation is.

Collocates such as *mauri* in (23) only occurred with the sense 'settle'. There were four examples where *tau* functioned as a post-posed modifier to *mauri*. Moorfield lists this as a multi-word unit meaning 'without panic' 'deliberate', This environment can again be used to tag for *tau* 'settle'.

23. …he ngākau māhaki, he mauri tau
…DET heart humble, DET emotions settle
'a placid heart equates to a harmonious state'.

An ambiguous example from the data was (24). This example is actually the 'song' sense in the sense of *tauparapara* 'chant'. This shows that the senses associated with *tau* which have similar grammatical functions will cause potential ambiguity in a corpus. The sense 'settle' could be mistaken as the correct sense here as *waka* 'canoe' is a frequent collocate of the sense 'settle'. However, reading back through the wider context, it is clear that the topic of discussion is the chant of *Mātaatua* and its meaning. The other sentence from this discussion containing *tau*, (25), is again an ambiguous nominal example outside of context. Another issue that (24) presents is that the modifying phrase *o te waka o Mātaatua* has the form of a possessive phrase which can be a subject in a nominalisation, therefore it has the appearance of an environment in which the verbal sense of *tau* could occur. Yet, the wider context gives the sense 'song'.

24.	koira	hoki	te	tau	o	te
	that is	PART	theSG	chant	of	theSG
	waka	o	Mātaatua			
	canoe	of	Mātaatua			

‘that indeed is the chant of Mātaatua’

25.	te	mauri	o	Mātaatua		
	the	essence	of	Mātaatua		
	kei	roto	i	taua	tau	rā
	PREP	inside	PREP	that	chant	there

‘the essence of Mātaatua is in that chant’

There were examples excluded from the analysis due to their ambiguity because the wider environment did not provide enough context to make the decision as to what sense of *tau* it was. Example (26) exemplifies one of these cases. The sense of *tau* here could be ‘be fitting’, but it could also be ‘sing’ functioning as a nominalisation. Examples like this from the MBC where the sense of the word was not clear-cut, were excluded from the analysis.

26.	…te	pai	hoki	o	ngā	tēpu,
	…theSG	good	INTENS	of	thePL	table
	te	tau	o	ngā	waiata	
	theSG	?	of	thePL	song	

‘the tables were well presented and the songs were sung’/

‘the tables were well presented and the songs were awesome’

There were only 29 instances of *tau* preceded by Ø phrase-marking in the MBC. All examples were the ‘settle’ sense except for example (27) which contained the sense ‘awesome’.

27.	Tau	kē	mai	te	pāti
	awesome	INTENS	hither	theSG	party

‘the party was awesome’

Overall the greatest indicators were the phrase-type markers. Where there were determiners preceding *tau*, this was a very high indicator for the sense *tau* ‘year’.

6.3 Conclusions

The results of the analysis provided the following set of rules that could be applied to tag a corpus for the lexemes *tau* ‘year’ and *tau* ‘settle’ which are by far the most frequent senses. The examples have been given in order of their likely reliability, based on the numbers of tokens involved, from greatest to least for each sense.

1. If a TAM precedes tau = ‘settle’ sense
2. Ø marking preceding tau = ‘settle’ sense
3. If det precedes tau = ‘year’ sense unless followed by directional particles, adverbial expressing goal or collocates associated with the ‘settle’ sense
4. If a cardinal number precedes or follows tau = ‘year’ sense
5. If hia precedes tau = ‘year’ sense
6. If ia precedes tau = ‘year’ sense
7. If rau precedes tau = ‘year’ sense
8. If an ordinal number follows tau = ‘year’ sense

7 Conclusion

The results showed that the lexemes from all three case studies could be identified in the corpus on the basis of consistent clues that occur in their linguistic environment. Assuming that my findings can be generalised, it is likely that if the adjacent syntactic parts of the phrase in which the lexeme occurs are examined, and the grammatical information supplied by the wider linguistic environment is taken into account, it would be possible to determine the appropriate lexemic tag for a word-form in a corpus in Māori.

The results from *kī* 'say' and 'full' showed that the adjacent elements in the phrase in which the word-form occurred can help to distinguish each lexeme. The most accurate indicator for the 'say' lexeme was the TAM *me* which only ever occurred preceding the 'say' lexeme. The directional particles *mai* and *atu* suggested the meaning 'say'; however the meaning 'full' was also found to co-occur with *atu*. The statistical probability of this though was very low and so it could be possible to tag using *atu*. An automated search could be made for most, if not all, of these features.

There were indicators outside of the phrase peripheries which it would not be possible to tag for using a computer program. For example, the most effective way of disambiguating the lexeme *kī* 'full' was to review the subject noun phrase that occurred following the verb constituent: if the subject noun phrase was an inanimate entity it was highly likely to be the 'full' lexeme. Word collocates falling into categories such as 'container' or 'vessel' would be more difficult to tag for than animacy or inanimacy as this would require entering all the possible collocates of 'full' into the computer program or marking every item in the lexicon with semantic features in enough detail to include this information. Even then, it is unlikely that the semantic features for a word like *mouth* would include anything to indicate that it was a container, although it can clearly be described as 'full'. Where there was no subject noun phrase, it was necessary to search the remainder of the syntactic construction. My results showed that an adverbial expressing cause, if there was one, was an important factor which indicated the sense 'full', but it would be very difficult for a computer to

distinguish a cause phrase from the many other possible phrase-types that can begin with *i* in Māori.

Due to the difference in syntactic function of the lexemes *mea* ‘say’ and *mea* ‘thing’, the phrase peripheries could be tagged to effectively distinguish each lexeme. The probability of this difference providing the appropriate answer is statistically high. The results could influenced by the spoken nature of the corpus, since the occurrences of *mea* ‘thing’ preceded by a TAM were hesitations and would less likely be found in a written corpus. Using the phrase periphery would not give the desired result when the lexeme *mea ‘say’* occurs in a stem nominalisation and is preceded by a determiner, though even then it was found that a directional particle would co-occur with *mea* ‘say’ in this environment and distinguish the appropriate meaning.

This is also the case for the lexemes *tau* ‘settle’ and *tau* ‘year’: their syntactic function makes it clear as to what lexeme we have in context. Where the verbal lexeme occurred in a stem nominalisation, again it was highly likely for a directional particle following the verb to be found to differentiate its meaning, and if there was no directional particle but an adverbial expressing goal was present, that also made it distinguishable. However, an adverbial expressing goal would need to manually tagged, because the preposition *ki* which marks goal phrases also has many other functions in Māori.

The common pattern across all three case studies is that the parts of the phrase are key indicators when tagging lexemes. It is not possible to tag for all clues that discriminate lexemes, such as collocates in the subject noun phrase and adverbials expressing cause or goal, as this goes beyond the scope of what is possible by today’s standard computer tagging programs. Manually generating answer keys (manually annotating for syntactic and semantic environments) for just *kī* ‘full’ would be time-consuming and tedious. Therefore these features would require manual annotation.

The patterns of these case studies are probably generalisable to other case studies on the assumption that language will not tolerate too great a burden of ambiguity, and if homonyms arose that frequently could not be distinguished by context, the language system would be likely to change in

some way to resolve the issue. However there is no guarantee that two homonymous verbal lexemes will pattern differently, but it is likely that there will be obvious syntactic clues to signal the meaning of each lexeme. Certain particles will differentiate most cases where there are two homonymous lexemes when one is nominal and one is verbal, although ambiguity is likely to occur where a verbal lexeme occurs in a stem nominalisation, but directional particles will often dictate a verbal interpretation.

The contribution my thesis makes to the issue of tagging corpora of Māori lies in its investigation of the most probable locations of the items that would be crucial for the discrimination of homonymous content lexemes. It developed a method for analysing contextual patterns for individual lexemes. The results point to the importance of the phrase periphery as the foremost location for clues. Because the items that occur in phrase peripheries in Māori fall into largely listable sets, it is possible to set up an automated search for them. This suggests that it should be possible to automate at least some part of the tagging process for Māori.

All that remains is to consider the areas for further research that are raised by my thesis. This thesis was only concerned with the analysis of content words as explained in 3.6.1, and therefore further investigation into the analysis of function words would provide a useful contrast here. The top ten most frequent words in Māori are function words as presented in Boyce (2006). There are possibly as many as eight or nine functions given in the review of *e* in section 3.6.2 and the most frequent uses of those functions are yet to be identified. In terms of tagging, it is not clear that all of the potential lexemes realised as *e* will be associated with distinct contextual clues, particularly the range of TAM uses. This mirrors the problem faced by a language learner, who looks up the form *e* in a dictionary, and finds as many as fifteen entries (as in Moorfield)! A dictionary is unlikely to provide the help a learner needs to determine the sense of *e* in a particular sentence.

Due to the issues surrounding the spoken nature of the MBC and its effect on the findings of this research, further investigation into the analysis of written material for high frequency words would be beneficial. The MBC has provided an account of high frequency words in spoken data and a comparable list for

high frequency words in written material would be a useful contrast.

John Cocks (personal communication) suggested using a bootstrapped model for annotating data in Māori. The first issue is that there have been few attempts at 'treebank' development. Bosco et al (2000:1) state:

> …treebank development involves an annotation process performed by a human annotator helped by an interactive parsing tool that builds incrementally syntactic representation of the sentence.

Ghayoomi (2012:1-2) states that computational approaches to tagging data are developed under human supervision in order to build as comprehensive a program as possible. This process is difficult, tedious and time consuming, resulting in these types of computer programs not being available for many languages. Ghayoomi investigates an alternative approach:

> Considering that a portion of the language is regular, we can define regular expressions as grammar rules to recognize the strings which match the regular expressions, and reduce the human effort to annotate further unseen data. In this paper, we propose an incremental bootstrapping approach via extracting grammar rules when no treebank is available in the first step

It is possible to build these types of programs for Māori but no means a simple task. Because there is little research into tagging Māori corpora, these approaches mentioned here are yet to be proven as effective for the Māori language. However, John Cocks (personal communication) mentioned that some programs could be viable for Māori but this is dependent on the types of resources one has at hand such as dictionaries and lexicons to speed up the process, of which there are few in comparison to English. Another area for research would be building the types of lexicons one needs in order to use some of the automated tagging programs available.

This thesis has attempted to answer a tiny portion of the questions involved in exploring the possibility of tagging corpora in Māori. The purpose of the case studies in this thesis was to investigate whether it is possible to determine which lexeme we have in any particular textual token. The thesis analysis provided a method for collecting patterns and showed it is possible in these cases to discriminate one from the other.

Mā whero, mā pango, ka oti .....'it is by red and by black that it is finished'

Bibliography

References

Aitchison, J., 1987. *Words in the Mind: An Introduction to the Mental Lexicon.* Oxford: Blackwell Publishers.

Atkins, B.T.S., 2002. Then and Now: Competence and Performance in 35 Years of Lexicology. In A. Braasch and C. Povlsen (eds). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002*. Copenhagen Center for Sprogteknologi, 1-28.

Atkinson, M., D. Kilby and I. Roca, 1982. *Foundations of General Linguistics*. London: Allen and Unwin.

Bauer, L., 1998. *Introducing Linguistic Morphology*. Edinburgh: Edinburgh University Press.

Bauer, W., W. Parker, T.K. Evans and T.A.N. Teepa, 1993. *Maori.* Descriptive Grammar Series. London: Routledge.

_____1997. *The Reed Reference Grammar of Māori.* Auckland: Reed.

_____2009. *Maori Vocabulary Size: Towards an explanation of Mary Boyce's findings.* Unpublished LALS paper presented at Victoria University, Wellington, 2009.

Bauer, L. & W. Bauer, 2012. The inflection-derivation divide in Māori and its implications*. Te Reo,* 55:3-24.

Belyayev, H., 1963. *The Psychology of Teaching Foreign Languages*. Oxford: Pergamon Press.

Biggs, B., 1969. *Let's Learn Māori: a Guide to the Study of the Māori Language.* Wellington: A.W. and A.H. Reed.

Bosco, C., V. Lombardo, D. Vassallo, L. Lesmo, 2000. Building a Treebank for Italian: a Data-driven Annotation Schema. [Electronic Paper]. *Proceedings of the Second International Conference on Language Resources & Evaluation, LREC 2000*. pp.1-7. [Accessed 12 June 2013.] Available from http://www.lrec-conf.org/proceedings/lrec2000/html/summary/220.htm)

Boyce, M.C., 2006. *A corpus of Modern Spoken Māori*. Unpublished PhD thesis, Victoria University of Wellington.

Chomsky, Noam, 1965. *Aspects of the Theory of Syntax*. Cambridge, Massachusetts: MIT Press.

Ghayoomi, Masood, 2012. From Grammar Rule Extraction to Treebanking: A Bootstrapping Approach. [Electronic Paper]. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012.* p.1-8 [Accessed 12 June 2013.] Available from: http://lrec-conf.org/proceedings/lrec2012/index.html

Haden, F.,1992. Frank Hadens' column. *Dominion Sunday Times*, July 26, 1992. Wellington: Dominion Newspaper.

Harlow, R., 2001. *A Māori Reference Grammar.* Auckland: Longman.

_____2007. *Māori: A linguistic Introduction*. Cambridge: Cambridge University Press.

_____and Peter Keegan, Jeanette King, Margaret Maclagan and Catherine Watson, 2009. The changing sound of the Māori language. In James N. Stanford and Dennis R. Preston (eds), *Variation in Indigenous Minority Languages.* Amsterdam: John Benjamins Publishing Company, pp.129-152.

Kilgarriff, Adam, 2008. I Don't Believe in Word Senses. In Thierry Fontenelle (ed.), *Practical Lexicography: A Reader*. New York: Oxford University Press, p.143-150.

Krupa, V., 1968. *The Māori Language.* Moscow: Nauka Publishing House Central Department of Oriental Literature.

_____1982. *The Polynesian Languages: Languages of Asia and Africa Volume 4.* London: Routledge & Kegan Paul Ltd.

Lyons, J., 1968. *Introduction to Theoretical Linguistics*. Cambridge: Cambridge University Press.

_____1977. *Semantics Volumes 1 & 2*. Cambridge: Cambridge University Press.

Nation, I.S.P., 1996. *Language Curriculum Design.* Wellington: English Language Institute Occasional Publication No. 16. School of Linguistics and Applied Language Studies, Victoria University of Wellington.

_____2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press..

Pawley, A., 1967. The relationships of Polynesian Outlier languages. *The* Journal *of the Polynesian Society*, 76(3):257-296.

Salmond, A., 1985. *Hui: A study of Māori Ceremonial Gatherings*. Auckland: Reed Methven.

Dictionaries

Biggs, B., 1990. *English-Maori Maori-English Dictionary.* Auckland University Press.

Greenhill, S.J. and R. Clark, 2011. POLLEX-Online: The Polynesian Lexicon Project Online. *Oceanic Linguistics*, 50(2):551-559.

Herbert W.W., 1957. *A Dictionary of the Maori Language* (2nd ed). Wellington: Government Printer.

Huia, 2006. *Tirohia Kimihia: A Māori Learner Dictionary.* Wellington: Huia Publishers.

Te Taura Whiri i te Reo Maori, 2008. *He Pātaka Kupu te kai a te rangatira.* Auckland: Penguin Group. [Accessed 5 August 2013.] Available from: http://www.tki.org.nz/tki-content/search?SearchText=%22tou+whiore%22&TKIGlobalSearch=1&x=-522&y=-282

Moorfield, J.C., 2005. Te Aka Māori – English, English – Māori Dictionary and Index. Auckland: Pearson.

Reed, A.W., 2001. *The Reed Concise Māori Dictionary.* Auckland: Reed Publishing (NZ) Ltd.

Ryan, P.M., 1995. *The Reed Dictionary of Modern Māori.* Auckland: GP Print Ltd, New Zealand.

Tregear, E., 1891. *Māori – Polynesian Comparative Dictionary*. Wellington: Lyon and Blair.

Williams, H.W., 1971. *A Dictionary of the Maori Language.* Wellington: Government Printer.

Appendix 1: Data Analysis of *kī*

The information contained in Appendix 1 is a selection of data from *kī* 'say'. The umlauts were used in Boyces (2006) MBC hence their use in the Concordance column.										
Concordance	Stem Nom	type	Sense	Prep	TAM	Det	Mod	Mod 2	Mod 3	Sentence/Clause Position
ahurihia e Kuru ana kauhau o mua atu, ka kï atu ki te rangatahi nei, e noho kout			say		Ka		atu			Predicate head
te nei ka hongi atu i a Te Kuru, anä, ka kï atu a, a Te Kuru ki a ia, anei te ka			say		Ka		atu			Predicate head
ou whakaaro me ngä kaumätua, në? Anä. Ka kï atu te, me kï rä, te tangata nei, te			say		Ka		atu			Predicate head
ngä kaumätua, në? Anä. Ka kï atu te, me kï rä, te tangata nei, te tangata Päkeh			say		me		rā			Predicate head
Päkehä nei a ki a ia, mehemea koe kei te kï mai ki ahau me haere mai te Pirimia			say		kei te		mai			Predicate head
nö a Te Kuru ki te körero ki a mätou, ka kï mai, i tana haramaitanga tuatahi i t			say		ka		mai			Predicate head
ka puta tonu atu ki waho rä anö. Anä, ka kï mai ia ki a mätou, i tënei wä tonu,			say		ka		mai			Predicate head
ia-tonu-nei, tahi rau paiheneti tä mätou kï atu ki a koutou inäianei, anä, hei t	Y		say			tā mātou	atu			Subject NP
, anä, hei täpiri atu ki wërä körero, ka kï mai ia ki a mätou, e noho koutou i k			say		ka		mai			Predicate head
tära ngä moni tohatohahia e ia, me tana kï atu anö, harikoa te rä ki a koe, kua	Y		say	me		tana	atu	anō		Prepositional phrase
a atu te kawenata o Aotearoa iäianei. Ka kï anö a ki te tutuki te whakaaro nei,			say		ka		anō			Predicate head
nei, i raro anö pea i ö rätou küare kua kï ngä pirihimana kei te whakateka atu			say		kua					Predicate head
hakapono ki ngä körero a te käwana. Ä, e kï ana anö a			say		e		ana	anō		Predicate head

te kei te haere tonu tënei										
Tämaki-makau-rau, täpiri atu ki tënei, e kï ana te ko ngä rïpoata e rapuhia nei			say		e		ana			Predicate head
tea mai i ngä mahi a räua ko. Nä reira e kï ana te ko ngä mahi a te käwana i tën			say		e		ana			Predicate head
o ngä mängai i tae atu ki tënei hui, kua kï , ka haere ake rätou, ä, ki Hämoa ki			say		kua					Predicate head
Tangaroa ko te Tari Pirihimana kë kei te kï he raruraru kei konei ka tü ana tëne			say		kei te					Predicate head
ätai atu ki a rätou, kei hea te körero e kï ana e kore rätou e ähei ki te hanga			say		e		ana			Predicate head
a ka pupü ake i roto i tënei kaupapa, me kï pea, ka haere atu ki roto i ngä tama			say		me					Predicate head
atu, nä reira i runga i tërä kaupapa me kï pënei pea, ko ä tätou taitamariki e			say		me					Predicate head
aenganui i ënei kamupene päkihi. Anä, ki kï anö a ko ngä iwi o täwähi e pupurihi			say		0		anō			Predicate head
u kia körero, engari, ko te tüpato koe e kï ake nei au, kia kaua e pöhëhëtia kei			say		e		ake	nei		Predicate head
nui i ngä mahi a ngä uri i Päkaitore. E kï ana a kua höhä ngä iwi o reira ki Te			say		e		ana			Predicate head
ia whakaotihia atu tënei mahi ä rätou. E kï ana ia, kei te rähui mai te tini me			say		e		ana			Predicate head
ö i runga i te taumata, moumou täima. Ka kï anö tënei uri, me hoki anö tätou ki			say		ka		anō			Predicate head
au. Nä rätou tënei whawhai, ä, ki täna e kï mai ana, kähore nä tëtahi atu iwi. M			say		e		mai	ana		Predicate head
hi. Ä, i rangona ai he aha a Ngäi Tahu i kï ai ko rätou kë e ähei ana hei kaitia			say		i		ai			Predicate head
a a Te Tiriti o Wai-tangi, ko tä rätou e kï ana, ko te tangata whenua kei a Käi			say		e		ana			Predicate head
a tätou, ngä mea ka whängaia, ka, ka, ka kï te käpata i			full		ka					Predicate head

te kai, në. Hei aha te.										
tou anö, hei aha, nä, te, tö rätou, ä, e kï ana, me haramai koutou ki ngä poukai			say		e		ana			Predicate head
mai rä anö i te tau iwa tekau mä tahi. E kï anö ana a, käre i tua atu i tënei ti			say		e		anō ana			Predicate head
iatangahia atu ki te kite i te täkuta. E kï ana te whaea o te tamaiti nei, i kar			say		e		ana			Predicate head
una atu he äporotï ki te whänau. Me tana kï anö, he ähua taumaha tonu ki te kimi	Y		say	me		tana	anō			Prepositional phrase
tea te mahi a, he whakaaro rangatira. Ka kï anö a Tau Henare, kei hea atu i tua			say		ka		anō			Predicate head
ia, kia riro mai rä anö i a tätou, ä, me kï , te kaiwhakahaeretanga o, o ënei tü			say		me					Predicate head
whakahaere hï ika Mäori, ko Matiu Rata e kï ana, käre he take o te hamuhamu haer			say		e		ana			Predicate head
te whare, ki reira torotoro ai, ä, ki te kï mai rätou, ä, hoki mai, ä, kei te mö	Y		say	ki		te	mai			Prepositional phrase
te möhio kei te pai tö mahi. Engari, ka kï mai rätou, mä mätou koe e waea atu,			say		ka		mai			Predicate head
mätauranga mehemea e pïrangi koe te, me kï pea te, te tono i ö, i ö taonga ki r			say		me		pea			Predicate head
tënei wä, ä, ki te katia aua höhipera e kï nei te käwana me kati, ki te kati au			say		e		nei			Predicate head
ränei koe i te, i te pouaka whakaata, e kï ana, ka whakahokia ngä uri ki ö räto			say		e		ana			Predicate head
, ki te i te käinga nei. Kaua, käre au e kï ana, heria mai ngä nëhi, ö, O waho,		neg	say		e		ana			Neg VC
a o te iwi! Kua hë hoki au i konei. Ä, e kï ana a roto i a au. Kei te mahi o, ët			say		e		ana			Predicate head
ki te tautoko i reira. Ä, mä rätou e, e kï mai, me haere pëhea tätou ki te taut		ae	say		e		mai			AE VC
te Tiriti o Wai-tangi. Tö tätou tiriti e kï nei, mä tätou			say		e		nei			Predicate head

anö tätou e whakahaere										
o ki a ia. He mea hou tënei, nö te mea e kï ana mätou te, te huarahi hei whakate			say		e		ana			Predicate head
tou te, te huarahi hei whakateretere, me kï pënei, ko ngä kairangahau a mätou me			say		me					Predicate head
tahi. Kia ora. Me mahi tahi. Kia taea te kï atu o tëtahi ki tëtahi, e whakaae an	Y		say			te	atu			Taea complement
tene tëtahi ki tëtahi. Në? Kätahi ka, me kï pënei, ka pakanga. I te mutunga kua			say		me					Predicate head