He Tātai Whenua: Automated Extraction of Landscape Terms and their Meanings in New Zealand Māori

K. Stock*[1], H. W. Morris[2], M. Forster[3], R. Paraku[4], and E. Egorova[1]

[1]Massey Geoinformatics Collaboratory, Massey University, Albany, New Zealand

[2]Department of Mãori Studies, Massey University, Palmerston North, New Zealand, hapu: Ngãi Te Rangitotohu, Ngãti Mãrau, Ngãti Maru

[3]Department of Mãori Studies, Massey University, Palmerston North, New Zealand, hapu: Rongomaiwãhine, Ngãti Kahungunu

[4]Department of Mãori Studies, Massey University, Palmerston North, New Zealand, hapu: Ngãti Tamaterã, Ngãti Maniapoto

*Email: k.stock@massey.ac.nz

Abstract

Current geographic information systems largely represent Western conceptualisations of landscape, and indigenous worldviews are rarely incorporated, resulting in decision making that does not consider all perspectives. The incorporation of New Zealand Māori worldviews in such systems requires a better understanding of the ways that Māori think about land. We use corpus linguistics to address this problem, demonstrating the automated extraction and analysis of content from early newspaper editions in the Māori language using word co-occurrences, which we then explore using an analytical frame developed for New Zealand Māori. We also demonstrate a method for automatically classifying geographic senses of landscape terms in the Māori using machine learning with a bag of words approach with SVM, a particular challenge due to the high degree of polysemy in the language.

Keywords: corpus linguistics, indigenous language, New Zealand Māori, landscape.

1 Introduction

Current geographical information systems (GIS) are largely reflective of Western conceptualisations of space and landscape, and fail to reflect many aspects of the worldviews of indigenous cultures. The inability of such systems to incorporate the deep, underlying perspectives of all stakeholders limits their effectiveness for decision making in geographical regions that are home to multiple cultural groups. In this paper, we bring together automated, text analysis approaches and social science, and demonstrate how text mining can be used to explore the semantics of specific landscape terms. Specifically, we apply corpus linguistics techniques to the investigation of descriptions of geographic feature types in New Zealand Māori, an Eastern Polynesian language from the Tahitic sub-group. While New Zealand Māori (hereafter references to Māori refer to New Zealand Māori

*

unless qualified) presents some challenges (e.g. high degree of polysemy; difficulties in part of speech delineation; few Natural Language Processing [NLP] tools), large numbers of texts in the Māori language are available, offering an opportunity that is not available for many other indigenous languages.

This paper analyses the use of a set of 10 selected landscape terms in New Zealand Māori, studying word frequency, co-occurrence and information gain, with the goal of gaining preliminary insights into the semantics of those terms, as part of a long term goal to study Māori conceptualisations of land as part of a broader understanding of Māori worldviews. Such analysis has not previously been performed on this language, and while some aspects of landscape descriptions have been studied both in New Zealand Māori (Murton, 2011, 2012) and in related languages (Cablitz, 2008), a corpus linguistics approach has not been applied. We also demonstrate the ability to predict the geographic sense of landscape terms in New Zealand Māori using machine learning, as a useful tool in building a corpus for more sophisticated text analysis in the future.

This research is part of He Tātai Whenua Te Ao Māori landscape classification project. This project brings together a team of indigenous and other researchers from a range of disciplines to synthesise a Te Ao Māori landscape classification that can be integrated with GIS to enable improved environmental reporting and monitoring that is cognisant of Māori worldviews and aspirations. It is important to mention from the outset that the design of this project, parts of the data gathering process and analysis have been guided by Māori language speakers, one of whom is a Senior Lecturer in the Māori language and licensed translator and interpreter. These members of the team have expert knowledge of the relationship between Māori and the land so are well placed to determine whether this analysis is reflective of Māori worldviews.

The contributions of the paper are three fold:

1. Firstly, we demonstrate a simple computational method for extracting semantics from a text corpus, in a particular domain for an indigenous language for which there are few available NLP tools. Previous study of indigenous conceptualisations, particularly in the landscape context, have mainly relied on field work (e.g. Mark et al. (2011)), while instead, we demonstrate how text resources can be used to extract useful knowledge.

2. Secondly, we provide preliminary insights regarding Māori conceptualisations of land, centred around 10 specific landscape terms. We map extracted co-occurring words to an analytical frame that reflects Māori worldviews to gain new insights into the way Māori see land.

3. Finally, we demonstrate a method for automatically classifying different senses of landscape terms, particularly to distinguish geographic from non-geographic senses, a task that is necessary for effective computational analysis of the semantics of landscape terms, due to the high incidence of polysemy in languages in the Eastern Polynesian group.

The paper is structured as follows. Firstly, we present background and related work. We then provide material about Māori worldviews, and present an analytical frame that is used later in the research. Following this, we describe the methodology used to extract and classify the data, and present the results of the analysis, including the classification of word co-occurrences and landscape term sense classification.

2 Background and Related Work

2.1 Critical GIS and ethnophysiography

GIS technology plays a key role in the decision-making processes in societies today. Given its ubiquity, it is of crucial importance to understand the intrinsic link between spatial representations embedded in GIS and power (Pickles, 1995). This concern triggered the emergence of the critical GIS approach in the 1990s and the recognition that concepts embedded in modern GIS represent largely Western worldviews and fail to account for indigenous conceptualizations (Rundstrom, 1995; Schuurman, 2000). GIScience has responded to these concerns by a renewed and deeper engagement with challenges such as communicating the meaning and defining semantics across cultural and language boundaries (Raubal et al., 2013). It has also acknowledged the necessity of an interdisciplinary cooperation with cognitive scientists and linguists, which was reflected in the emergence of ethnophysiography (Mark et al., 2011) and landscape ethnoecology (Johnson, 2010; Johnson and Hunn, 2010), both focussed on the cross-cultural study of landscape categories. This line of research mostly explores landscape conceptualizations in particular indigenous communities such as Yindjibarndi in Australia and Navajo in the US (Mark et al., 2011), or compares landscape categorization in culturally and linguistically different communities that occupy ecologically similar natural environments (Holton, 2011). Methods are essentially qualitative and are based on fieldwork, including interviews and other forms of elicitations, such as photo descriptions or sketches. Our work differs from these previous approaches in that we study indigenous geographic conceptualisations using computational text analysis and mining.

2.2 Māori language

New Zealand Māori, also known as te reo, is the language of the Māori people of New Zealand. Due to an aggressive suppression policy introduced by British colonisation (Waitangi Tribunal, 1986), today only 50,000 adults of mostly Māori descent (11% of the Māori population) report a capacity to speak te reo well or very well (Statistics New Zealand, 2013), but significant efforts are underway to revitalise the language. New Zealand Māori "belongs to the Polynesian subgroup of the huge Austronesian language family, which consists of over 700 languages" (Harlow, 2007, p. 1). Most specifically, it is a member of the Tahitic language partition, which also contains Tahitian, Rarotongan, Tuamotuan, and is itself a member of the larger Eastern Polynesian group (Pawley, 1966).

Māori conceptualisations of landscape differ from a Western perspective (from the self-centred perspective, gazing upon from a distance). This difference is attributed to a worldview that emphasises the spiritual dimension and kinship relationships between people and land (Roberts et al., 1995; Royal, 2003). Māori place names and names of geographic features are derived from a close association with the natural world and reflect a need to orally record the histories and connections of a people to place to establish tribal authority. For example, many place names refer to events that occurred in the place (indicated by place names with prefix o), but many names now are shortened or abbreviated from these longer, event-describing names, and incorporate geographic feature types, some of which are also polysemous with body parts or combine terms associated with both the land and body (e.g. Te Mata o Rongokako – The Face of Rongokako, a ridge). Māori place names are also commonly inter-related, with groups of names sometimes describing the footsteps of the ancestors as they took a journey, or some other story or event (Davis et al., 1990; Murton, 2011; Hakopa,

2011). Polysemy is common in Māori, and multiple interpretations may sometimes be made of the meaning of place names when broken into component parts in different ways, with knowledge of the history of a place being needed in many cases for correct interpretation. Work on another member of the Eastern Polynesian language group, Marquesan, has highlighted challenges such as frequency of polysemy; the difficulty in identifying word classes (parts of speech), particularly the distinction between verbs and nouns; and the tendency for place names to incorporate landscape terms, many of which also apply to New Zealand Māori (Cablitz, 2008).

2.3 Maori language analysis tools

Corpus linguistics analyses of New Zealand Māori and other indigenous languages, particularly in the geographic domain, are very limited. Furthermore, tools that could assist with such work with New Zealand Māori are scarce, being limited to the description of a formalised grammar (Bayard et al., 2002) and development of a tool for sentence parsing and generation which is no longer available (Knott et al., 2001, 2002, 2003); the development of tools for speech recognition (Bagnall et al., 2017); a diacritic restoration approach for New Zealand Māori (Cocks and Keegan, 2011); a method for identifying parallel text in English and Māori corpora (Mohaghegh and Sarrafzadeh, 2016) and a very detailed part of speech tag set for Māori (Cocks, 2012). Alternative approaches to common statistical approaches for minority languages have been proposed (Streiter and De Luca, 2003), and work on developing POS taggers for the related Cook Islands Māori (Coto-Solano et al., 2018) may be helpful in bridging the gap.

While this previous research shows related activities in a number of directions, our work focuses particularly on automated, corpus linguistics approaches to the study of landscape terms in New Zealand Māori, and we use this approach to better understand conceptualisations of land through language use.

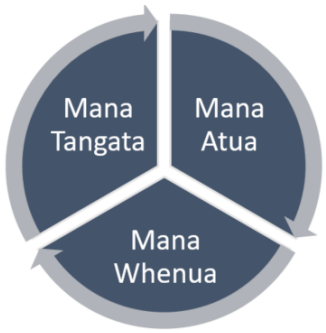
3 The Māori Worldview

Māori conceptualisations of space and landscape are understood through whakapapa/genealogy and kōrero tuku iho/genealogical narratives. According to this view all life is interconnected, traced back to the primodial parents – Ranginui/Sky father and Papatūānuku/Earth Mother (Buck, 1925; Walker, 1990). It is their children who populated the Earth with natural resources, flora and fauna – all the essential elements for survival and prosperity of humanity (Best, 1976, 1982).

Humanity were formed from the earth, the body of Papatūānuku, and imbued with the essence or key characteristics of her children (Smith, 1915). Māori people are known as tangata whenua meaning people of or from the Earth. Tangata whenua therefore is a reference to kinship shared with the primodial family and the environment and a reminder that tribal life and survival was dependent on a close association with the natural world. This thinking permeates a Māori worldview establishing norms and values that regulate behaviour (Roberts et al., 1995; Royal, 2003) and finds expression through the customs and the Māori language.

Place names and the names of geographic features provide information about the land, about the relationships of people with the land and/or resources and about place. Some names reference Māori origin narratives or conceptualise the land as a body. Others, record some aspect of tribal history and identity and, a large number describe the physical features of a landscape – the terrain

and resources of an area (Davis et al., 1990). Māori names associated with the landscape typically "emphasis spiritual values of land and provide the basis of tribal identity" (Davis et al., 1990, p. 8). To this end a simple analytical frame (Durie, 1998) that acknowledges these dimensions is used in this research to guide analysis of geographic feature types and determine validity of the automated analysis scheme. The analytical frame acknowledges the mana and mauri (or authority and life force) of the three key actors in the Māori worldview outlined above. The first set of actors are the atua or spiritual elements represented by the primodial family – Ranginui, Papatūānuku and their children. The second set of actors are the natural resources and flora and fauna, those tangible elements associated with whenua/land. The final set of actors are tangata/people. Figure 1 is a visual representation of the analytical frame summarising how it is applied to an analysis of geographic feature types. This analytical frame ensures that interpretation of data is grounded within a Māori worldview so that any inferences make sense from a cultural view point. This approach is consistent with a Māori political agenda that seeks to ensure that research reaffirms indigenous knowledge and ways of knowing and facilitates transformative change for Māori communities (Smith, 1999).



Mana atua: Place names and geographic feature types associated with spiritual elements of the environment. This dimension provides the source of mana and mauri.

Mana whenua: Place names and geographic feature types that describe the landscape from a Māori worldview.

Mana tangata: Place names and geographic feature types associated with ancestors and ancestral events that establish tribal authority and identity over specific geographical spaces and natural resources.

Figure 1: Analytical Frame

4 Method

Class	Meaning
p	A place name. e.g.The marae is by the Manawatu River."
f	A general landscape feature. e.g. "Rivers are common food sources."
s	A specific instance of the feature. e.g. "It is the river where we gather food"
n	A person's name which refers to a specific geographic feature, as it is common for people to be named after a feature to which they are connected in some way
r	A related landscape term. e.g. waipuke means to flood but is literally translated as water hill.
0	All non-geographic senses of the word

Table 1: Landscape Term Classes

For the purposes of this research, we extracted a sample of a Māori magazine and a Māori language newspaper from existing digital archives, being the first 20 issues of Te Ao Hou, covering the period

code	awa	kāinga	maunga	motu	puke	puna	repo	roto	wāhi	whenua	total
p	0.2%	0.0%	0.0%	0.5%	0.0%	1.2%	0.0%	0.0%	0.0%	0.2%	0.2%
f	1.9%	48.0%	33.3%	27.6%	18.7%	10.4%	5.5%	1.1%	17.2%	63.4%	12.7%
s	2.2%	24.3%	56.5%	40.0%	13.3%	1.2%	2.7%	0.2%	7.5%	12.6%	6.5%
n	0.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.1%	0.0%	0.2%	0.2%
r	1.8%	3.0%	7.2%	2.4%	37.3%	3.6%	0.0%	0.0%	2.3%	14.1%	3.3%
non-geo	93.5%	24.8%	2.9%	29.5%	30.7%	83.5%	91.8%	98.6%	73.0%	9.5%	77.2%
total geo	163	152	67	148	52	41	12	13	184	430	1262

Table 2: Percentage of Word Classes in the Corpus

from 1952 to 1957 inclusive[1] and the first 10 issues from Te Puke Ki Hikurangi, [2] all from 1897. Te Ao Hou was first published by the then Māori Affairs Department in 1952 and continued until 1976. Its aim was to act like a marae (meeting ground), and it featured articles written in both the English and Māori languages (Curnow, 2002). Te Puke ki Hikurangi was one of 34 Māori newspapers published between 1897 and 1913, and was the official newspaper of the Te Kotahitanga – Māori Parliament (Curnow, 2002). It was run exclusively by Māori. In both cases, the earliest available issues were used, in order to reduce the influence of English language knowledge from the writers of the newspapers concerned. The selection of the Māori newspaper Te Puke ki Hikurangi as one of the sources to extract files was based on the knowledge that the newspaper was an independent Māori newspaper edited and published by Māori. A text version of each of the selected editions was obtained and used for the analysis. English language content was ignored for the purposes of the analysis (other than as discussed in Section 5.3 when it proves useful for discriminating relevant content).

This corpus in total consisted of 794,649 word tokens and 29,837 word types. Ten commonly used landscape terms were selected from a larger set, based on the frequency of occurrence of those words in the corpus through a manual process with initial selection of candidate terms by the Māori speaking co-authors and then confirmation based on the combination of frequency within the corpus and examination of the context of the selected landscape terms, as while some candidate landscape terms had a substantial number of occurrences, these were not always with landscape senses. Following this process, the finally selected terms were: awa/river, channel; maunga/mountain; motu/island; puke/hill; repo/swamp, marsh; puna/spring; roto/lake; whenua/land; wāhi/place and kāinga/home. All instances of the terms were extracted from the newspaper editions using AntConc [3], a corpus analysis toolkit, including both the use of the terms as part of a larger word (which is common due to the high frequency of compound words in Māori) and the term as an independent word. These were then manually classified to identify uses of the terms in the landscape sense using the scheme in Table 1.

5 Analysis of Landscape Terms in New Zealand Māori

0.16% of the words in corpus are landscape senses of these ten terms. As can be seen in Table 2, the terms vary in the proportion of non-geographic use. Unsurprisingly, this is correlated with the length of the term, as shorter terms are more likely to form parts of other words with different senses, unlike

[1] http://teaohou.natlib.govt.nz/journals/teaohou/allthumbnails.htm
[2] http://www.nzdl.org/cgi-bin/library?a=p&p=about&c=niupepa
[3] http://www.laurenceanthony.net/software/antconc/

longer terms like maunga and whenua. The frequency of related words for some terms is mainly related to the presence of a longer word that contains the term. For example, whenua has a number of related senses including ahuwhenua (cultivated) and tuawhenua (mainland, inland).

Classes f (general landscape feature) and s (reference to a specific, but unnamed, landscape feature) were the most frequently represented, with surprisingly few p (place names) appearing. There are a large number of Māori place names that include the selected geographic terms (e.g. Maunganui, Motutapu, Rotorua), but these were not frequently contained in the editions selected for the analysis. While the selected editions do frequently contain place names, those containing the selected landscape terms seem to appear only infrequently in the limited number of editions selected for this analysis. In future work on a larger corpus of editions and wider set of sources, larger representation of place names containing these landscape terms may be expected.

5.1 Term Co-occurrence

In order to focus only on landscape senses of the selected terms, we created two sub-corpora from the original corpus: each containing text fragments (50 characters on either side of the landscape term), one with only geographic senses of the landscape terms, and one with only f (general geographic feature) senses, this being the most frequent class. From the original ten terms, five were selected for co-occurrence analysis on the basis of data volume, being those for which the selected term was used with a landscape sense more than 100 times across the corpus in total. While larger term frequency would have been desirable, the lists of co-occurring words were stable. For each of the five highly occurring terms (awa, kāinga, motu, wāhi, whenua), we extracted the top ten immediately preceding words (1L); the top ten immediately succeeding words (1R) and the top ten words in a 3L to 3R window. The top ten words were determined by Mutual Information, which is a measure of probability of a term occurring near the co-occurring word, adjusted for frequency across the corpus. In addition to these term specific lists, we extracted keywords from each of the two subcorpora, using the entire, original corpus (containing all non-landscape senses of the terms) as a reference for calculating Keyness. Table 3 classifies the words appearing in the extracted lists using the key cultural concepts from Section 3. The classification was performed manually and involves some subjective judgement.

The interactive verbs are interesting in that they vary between the different geographic features, both in quantity and nature. Motu and awa have far fewer of these verbs than the other features, and the verbs associated with kāinga are notable in the types of interactions they describe, being verbs of occupation, affection, care and connection. Those related to motu, awa and whenua are much more related to physical activities and management of the land, while those co-occurring with wāhi represent a range of types of relationship, perhaps indicating the generality of wāhi as a concept, potentially referring to a number of different types of places. Perhaps unexpectedly, words indicating possession are relatively rare, except for kāinga. While the stronger sense of possession/belonging for kāinga is not suprising, the clear evidence of close relationships between Māori and the landscape (for example, in self introductions, it is important for Māori to acknowledge the awa/river and maunga/mountain to which they belong and have a spiritual connection as part of their identity) might have led us to predict that verbs of possession would also be present for other types of geographic features, but while they are present, they are infrequent in comparison to kāinga. In future research we will further consider this aspect, and investigate whether this sense of belonging might be conveyed in language in more nuanced ways that this methodology has not identified. Also unexpectedly, references to body parts are rare in the data, and knowledge of Māori worldviews

would suggest a higher frequency of mention of these terms. These kinds of terms are more common in mōteatea/ancient chants and pūrākau/narratives that include whakapapa/genealogy.

landscape term	key cultural concept	n	explanation
awa/river	Mana Atua	1	One word indicating a body part (ngutu/mouth) co-occurs with awa, reflecting the Māori conceptualisation of the land as a person.
	Mana Whenua	15	Words for other geographic features were common among co-occurring terms, including roto/lake; repo/swamp; keri/hole; awaawa/valley. This was particularly true of the words containing awa, 4 out of 6 of which were landscape features
		5	A number of words provide more detailed descriptions of the landscape, expressing nuances in landscape term meaning, including words related to size (ririki/small, soft; nunui/big), and other descriptive characteristics (tūpā/barren).
		2	Co-occurrences that indicate possession (tōku/my) appear, expressing the close links between land and Māori sense of identity, and the worldview that Māori belong to the land.
	Mana Tangata	4	A number of co-occurring verbs indicate direct interaction between people and the land (paretai/to scrape soil; toremi/to drown, submerge), indicating the connection between identity and belonging, due to events and actions that occur in a place.
kāinga/home	Mana Whenua	3	Words for geographic features relate mainly to built features, including marae/courtyard; whare/house; rua/pit.
		4	Descriptive words co-occurring with kāinga include hou/new; tūturu/permanent, authentic and waimaria/lucky.
		10	Possessive words are numerous for kāinga, particularly in the 1L position, and include tōna/his, her; koutou/their; tōu/your; mātou/our and whiwhi/to have.
	Mana Tangata	11	Verbs indicating interaction with kāinga include noho/to sit, settle, occupy; whitiki/to bind; tiaki/to look after and tōrere/to desire, with these verbs particularly representing a close and intimate relationship with kāinga.
motu/island	Mana Atua	1	Upoko/head co-occurs with motu, suggesting a conceptualisation of the land as a body.
	Mana Whenua	4	Co-occurrence of geographic feature words are again much less frequent than for awa, referring to kauri (a kind of tree); toka/rock and waiawa/river.
		3	Descriptive words that co-occur with motu include tauhou/strange, exotic, unfamiliar and ririki/small, soft.
		2	Also less frequent are words indicating possession, which include tāua/you and me exclusive and nāku/belonging to me.
	Mana Tangata	6	Verbs that indicate close interaction are less frequent than for kāinga, and less intimate, including tatari/to wait (especially in the 1L position); whakarongo/to listen; whakanoa/to remove restriction and tōpū/to consolidate, combine.

Continuation of Table 3

landscape term	**key cultural concept**	**n**	**explanation**
wāhi/place	Mana Whenua	3	As for kāinga, the number of co-occurring geographic feature words was relatively small and biased towards built features, including pari/cliff; nohoanga/seat, dwelling and wātea/open space.
		4	Descriptive words that appear with wāhi include matatea/open, clear, free; hōhonu/deep and teitei/tall.
		2	Possessive words were relatively rare for wāhi, including tāua/our and o ngā /of the..
	Mana Tangata	15	Verbs indicating interaction include hanga/to build (in 1L position); whakarihariha/to be disgusted; huihui to come together (both in 1R position) and other verbs such as whiriwhiri/to choose; whakawehe/to divide; whakaeke/to attack, invade and whakatapu/to place a tapu on.
whenua/land	Mana Whenua	4	Geographic feature words that co-occur with whenua include tuawhenua/mainland, inland (containing word); ahu/mound; takiwā/district, place and kāuru/head of river or tree.
		3	Descriptive words that suggest a particular kind of relationship with the land include ngāwari/easy and ahuwhenua/cultivated.
	Mana Tangata	15	Interactive verbs that co-occur with whenua include whakawehe/to divide; whakawhāiti/to compress; whakawairākau/to fertilize, nourish; horomia/to swallow and whakatōpū/to combine. These are verbs that suggest management of the land in various ways.

Table 3: Classification of types of co-occurring words by key cultural concepts. The column headed **n** indicates the number of times a term appears in one of the lists of co-occurring words (but each appearance may represent multiple occurrences of the word on each of the lists).

Analysis of the top 50 keywords across each of the two sub-corpora in their entirety showed that geographic features occurred frequently, but this was largely due to the method for creating the sub-corpora, which resulted in larger numbers of the ten selected landscape terms. The next most numerous classes of words were the interactive verbs (6 in total across both lists), including the verbs haere/to go; hoki/to return; noho/to sit, settle and taha/to pass on the side, and the possessive words (both classes have 6 appearances in total across both lists). The majority of the keywords were particles, which are very frequent and numerous in Māori, many with spatial relationship senses (e.g. hei/at, on, in, and kei/at, on, in).

5.2 Classification

In a second investigation, we looked at the data in another way by performing a machine learning classification using a Support Vector Machine (SVM) classifier with a bag of words approach on the entire corpus. The purpose of this was twofold: (1) to evaluate the success of this method in classifying the text, as our future plans include the creation of a much larger corpus, and the ability to automatically classify extracted text will make this task more practical; and (2) to study the contribution of different words in terms of information gain for the classification task.

The process involved creating a document-word matrix, with each document being an instance of one of the ten selected landscape terms plus a window of ten words on either side, from the corpus described in Section 4; and each word being the most frequently occurring 1000 words across the corpus (i.e. the 'bag of words'). The cells were populated with the term frequency-inverse document frequency (tf-idf) value for the document and the word concerned. Thus each word in the bag of words became a feature in the model that was used by the SVM classifier, along with the manually annotated classes from Table 1 which were used for training and evaluation, with ten fold cross validation. The precision, recall and f-measure achieved for binary and multiclass classifications are shown in Table 4.

This suggests that automatic classification on a binary basis is possible, and a larger training set could be expected to improve these figures further. Automated classification into the 6 classes would require a substantially larger training set. Analysis of the attribute information gain for each of the words in the bag of words matrix unsurprisingly identifies the ten selected landscape terms as strong class predictors, representing 14 of the 30 words in the combined lists of the 10 words with the highest information gain from each of the three classifications in Table 4. 11 of the 30 words in the same list were English words, as these are strong predictors of the non-geographic use of the ten landscape terms. Since the classes in Table 1 other than 0 only contains Māori words, all English sentences in the analysis were classified as 0. This mainly occurs when the landscape terms appear as part of another word (e.g. repo as part of report). The remaining words in the list of 30 were common Māori words, including te/the (singular) and nga/the (plural).

Classification	**Precision**	**Recall**	**F-Measure**
Classification using all 6 classes – weighted average (some classes with low numbers have lower figures).	0.843	0.847	0.843
Binary classification into 0 vs any other class (any type of geographic use of the terms).	0.911	0.911	0.911
Binary classification into f (use of the term as a geographic feature type) vs any other class.	0.903	0.905	0.904

Table 4: Bag of Words SVM Classifier Results

6 Conclusions and Future Work

This paper has described an exploratory study that applies corpus linguistics techniques to understand and describe Māori conceptualisations that relate to specific landscape terms, showing that such approaches can provide useful insights in these kinds of studies, and that automated classification of landscape senses can be successfully achieved at the binary level.

While we considered the authenticity of 'the Māori voice' in selecting our language sources, in particular incorporating Te Puke ki Hikurangi as a publication written and managed by Māori; newspapers are nevertheless a medium for mass communication, and as such are likely to present a particular, common perspective. By its nature, this source does not detect differences between individual Māori world views, or those of particular groups (e.g. Māori women). Furthermore, the method we have demonstrated here takes an aggregated approach, in that the views across the corpus were considered as a whole. Our focus in this work has been to consider more general, commonly held perspectives, but corpus linguistics approaches could also be applied to study differences among the world views of groups and individuals. In future work, it is our intention to include a broader range of data sources, creating a much larger corpus that incorporates a wider range of Māori language publications, including reports from the Waitangi Tribunal Court process, as well as words from songs and chants and potentially more individual language sources like letters and journals.

We then plan to perform much more sophisticated analysis to gain a deeper understanding of Māori conceptualisations of land, and the ways in which they are expressed in language. In addition, NLP tools for the Māori language are required to enable some of these kinds of analysis, and we hope to apply existing work to achieve this end. Our future work will also address the issue of diacritics, which are absent from the publications that were used for this research, but that nevertheless are necessary for understanding of Māori, as in some cases the presence of a diacritic changes the meaning of a word. Current efforts to revitalise the Māori language are encouraging the use of diacritics, and we will explore the use of automated diacritic restoration algorithms to assist in this task.

7 Acknowledgements

This work was funded by a New Zealand Ministry for Business, Industry and Innovation (MBIE) Research Programme: He Tātai Whenua. The comments of the anonymous reviewers are gratefully acknowledged.

8 References

Bagnall, D., K. Mahelona, and L.-J. Peter
2017. Korero Maori: a serious attempt at speech recognition for te reo Maori. Auckland.

Bayard, I., A. Knott, and J. Moorfield
2002. Syntax and semantics for sentence processing in English and Maori. In *Proceedings of the 2nd Australasian Natural Language Processing Workshop', Canberra Australia*, Pp. 33–40. Citeseer.

Best, E.
1976. *Maori Religion and Mythology Part 1 | NZETC.* Wellington, New Zealand: Government Printer.

Best, E.
1982. *Maori Religion and Mythology Part 2 | NZETC.* Wellington, New Zealand: Government Printer.

Buck, T.
1925. *The coming of the Maori.* Nelson, NZ: Cawthron Institute.

Cablitz, G. H.
2008. When "what" is "where": A linguistic analysis of landscape terms, place names and body part terms in Marquesan (Oceanic, French Polynesia). *Language Sciences*, 30(2-3):200–226.

Cocks, J.
2012. Diacritic Restoration and the Development of a Part-of-Speech Tagset for the Māori Language. Master's thesis, University of Waikato.

Cocks, J. and T. T. Keegan
2011. A word-based approach for diacritic restoration in maori. In *Proceedings of the Australasian Language Technology Association Workshop 2011*, Pp. 126–130.

Coto-Solano, R., S. A. Nicholas, and S. Wray
2018. Development of Natural Language Processing Tools for Cook Islands Māori. Pp. 26–33.

Curnow, J.
2002. A brief history of Maori-language newspapers. In *Rere atu, taku manu! Discovering history, language and politics in the Māori language newspapers. Jenifer Curnow, Ngapare Hopa & Jane McRae (eds.)*, Pp. 17–41. Auckland, New Zealand: Auckland University Press.

Davis, T. A., T. O'Regan, and J. Wilson
1990. *Nga Tohu Pumahara: The Survey Pegs of the Past. Understanding Mâori Place Names.* Wellington, New Zealand: New Zealand Geographic Board.

Durie, M.
1998. *Te Mana, Te Kāwanatanga: the politics of self determination.* Auckland, New Zealand: Oxford University Press.

Hakopa, H.
2011. *The paepae: spatial information technologies and the geography of narratives.* Thesis, University of Otago.

Harlow, R.
2007. *Maori: A Linguistic Introduction.* Cambridge University Press. Google-Books-ID: RkJTxN-bcv7oC.

Holton, G.
2011. Differing conceptualizations of the same landscape. *Landscape in language: Transdisciplinary perspectives*, 4:225.

Johnson, L. M.
2010. *Trail of Story, Traveller's Path: Reflections on Ethnoecology and Landscape.* Athabasca University Press.

Johnson, L. M. and E. S. Hunn
2010. *Landscape Ethnoecology: Concepts of Biotic and Physical Space.* Berghahn Books.

Knott, A., I. Bayard, S. de Jager, L. Smith, and J. Moorfield
2001. A question-answering system for English and Maori. In *Proceedings of the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems (ANNES)*, Pp. 223–228, Dunedin, New Zealand.

Knott, A., I. Bayard, S. De Jager, and N. Wright
2002. An architecture for bilingual and bidirectional nlp. In *Proceedings of the 2nd Australasian Natural Language Processing Workshop (ANLP 2002)*. Citeseer.

Knott, A., J. Moorfield, T. Meaney, and L.-L. Ng
2003. A human-computer dialogue system for Maori language learning. Pp. 3336–3343. Association for the Advancement of Computing in Education (AACE).

Mark, D. M., A. G. Turk, N. Burenhult, and D. Stea
2011. *Landscape in Language: Transdisciplinary perspectives*. John Benjamins Publishing. Google-Books-ID: qcJxAAAAQBAJ.

Mohaghegh, M. and A. Sarrafzadeh
2016. Parallel Text Identification Using Lexical and Corpus Features for the English-Maori Language Pair. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Pp. 910–915.

Murton, B.
2011. Embedded in place: 'Mirror knowledge' and 'simultaneous landscapes' among Maori. In *Landscape in language: Transdisciplinary perspectives*. Benjamins.

Murton, B.
2012. Being in the place world: toward a Māori "geographical self". *Journal of Cultural Geography*, 29(1):87–104.

Pawley, A.
1966. Polynesian Languages: A Subgrouping Based On Shared Innovations In Morphology. *Journal of the Polynesian Society:*, 75:39–64.

Pickles, J.
1995. *Ground truth: The social implications of geographic information systems*. Guilford Press.

Raubal, M., D. Mark, and A. Frank
2013. *Cognitive and linguistic aspects of geographic space - New Perspectives on Geographic Information Research*. Berlin: Springer.

Roberts, M., W. Norman, N. Minhinnick, D. Wihongi, and C. Kirkwood
1995. Kaitiakitanga: Maori perspectives on conservation. *Pacific Conservation Biology*, 2(1):7–20.

Royal, C. e.
2003. *The woven universe: selected writings of Rev. Māori Marsden*. Otaki, New Zealand: Estate of Rev. Māori Marsden.

Rundstrom, R. A.
1995. Gis, indigenous peoples, and epistemological diversity. *Cartography and geographic information systems*, 22(1):45–57.

Schuurman, N.
2000. Trouble in the heartland: Gis and its critics in the 1990s. *Progress in human geography*, 24(4):569–590.

Smith, L. T.
1999. *Decolonizing Methodologies: Research and Indigenous Peoples.* Dunedin, New Zealand: Zed Books Ltd. Google-Books-ID: 8R1jDgAAQBAJ.

Smith, P.
1915. *The lore of the whare wananga. Written by H.T. Whatahoro from the teachings of Te Matorohanga and Nepia Pohuhu. Translated by Percy Smith.* New Plymouth, New Zealand: [publisher unknown]. Google-Books-ID: mvmUHHyt99UC.

Statistics New Zealand
2013. *Te Kupenga 2013 (English) | Stats NZ.*

Streiter, O. and E. De Luca
2003. Example-based NLP for minority languages: tasks, resources and tools. In *Proceedings of the Workshop "Traitement automatique des langues minoritaires et des petites langues", 10e conference TALN.*, Pp. 233–242, Batz-sur-Mer, France.

Waitangi Tribunal
1986. *Report of the Waitangi Tribunal on the te reo Maori claim (Wai 11).* Wellington, N.Z.: Brookers. OCLC: 946525712.

Walker, R.
1990. *Ka whawhai tonu matou: Struggle without end.* Auckland, New Zealand: Harmondsworth: Penguin.