Māori Loanwords: A Corpus of New Zealand English Tweets

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN
the ROYAL SOCIETY of NEW ZEALAND
TE APĀRANGI

David Trye
Department of Computer Science
University of Waikato, New Zealand
dgt12@students.waikato.ac.nz

Andreea S. Calude
School of General and Applied Linguistics
University of Waikato, New Zealand
andreea.calude@waikato.ac.nz

Felipe Bravo-Marquez
Department of Computer Science
University of Chile & IMFD, Chile
fbravo@dcc.uchile.cl

Te Taka Keegan
Department of Computer Science
University of Waikato, New Zealand
tetaka.keegan@waikato.ac.nz

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato
KO TE TANGATA

Method

Collect Tweets

We used the *Twitter Search API* to harvest 8 million English-language tweets containing one or more Māori words (loanwords) from a predefined list. These tweets were used to create the *Raw Corpus*.

Label Samples

Proud to be a **kiwi** ✓
Love my crazy **whānau** ✓
Moana is my fav princess ✗
haka ne kuma fa you say ✗

We extracted a random sample of tweets for each query word and labelled these as "relevant" or "irrelevant", depending on the context. The annotated tweets, which comprise the *Labelled Corpus*, became our training data (after removing all query words that were irrelevant more than 90% of the time).

Transform Dataset

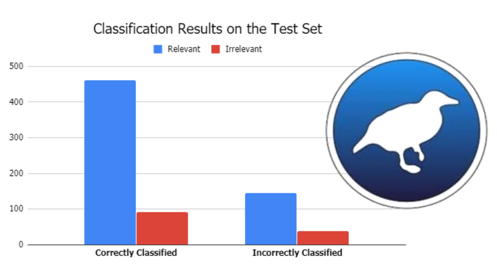
Vocab
$word_1$
$word_2$
...
$word_n$

Tweet Vectors

	$word_1$	$word_2$	...	$word_n$
$tweet_1$	0	1	...	0
$tweet_2$	1	0	...	1
...	...	...	...	...
$tweet_m$	1	1	...	0

We converted our data into a suitable format for machine learning. This involved transforming the tweets into vectors based on the word n-grams they contain.

Build Classifier



Using stratified, independent test and training sets in Weka, we experimented with various machine learning models, including *Naive Bayes Multinomial* and *Linear Logistic Regression* (with different word n-grams). We evaluated the model with the best performance.

Deploy Model

Raw Corpus
$tweet_1$ ✓
$tweet_2$ ✗
$tweet_3$ ✓
...
$tweet_m$ ✓

We deployed this model on the *Raw Corpus* to obtain automatic predictions for the relevance of each tweet. We then removed all tweets that were classified as irrelevant (p<0.5), thereby producing the *Processed MLT Corpus*.

Introduction

The indigenous language of New Zealand is **Māori**, spoken by roughly 4% of the New Zealand population. Māori is an Austronesian language which constitutes the last stop on the "island-hopping train" originating in Taiwan.

Map of the world, highlighting New Zealand's location. Source: Wikimedia

Words that are borrowed from another language are called loanwords. **Māori loanwords** are widely used in New Zealand English (NZE) for various social functions by New Zealanders within and outside of the Māori community. Motivated by the lack of linguistic resources for studying how Māori loanwords are used in social media, we present a new **corpus of New Zealand English tweets**.

We collected tweets containing selected Māori words that are likely to be known by New Zealanders who do not speak Māori. Since over 30% of these words turned out to be irrelevant (e.g. *mana* is a popular gaming term; *Moana* is a character from a Disney movie), we manually annotated a sample of our tweets into relevant and irrelevant categories. This data was used to train **machine learning models** to automatically **filter out irrelevant tweets**.

New Zealand English

One of the most salient features of New Zealand English (NZE) is the widespread use of Māori words (loanwords), such as **aroha** (love), **kai** (food) and **Aotearoa** (New Zealand). Below are four examples of real tweets containing a rich variety of loanwords (emphasised in **blue**):

(1) Sorry I thought you were **Kiwi** [a New Zealander]. **Aotearoa** is the **Māori** name for NZ (1064121983678406656)

(2) Led the **waiata** [song] for the **manuhiri** [guest] at the **pōwhiri** [welcome ceremony] for new staff for induction week. Was told by the **kaumātua** [elder] I did it with **mana** [pride] and integrity. (757369343642480640)

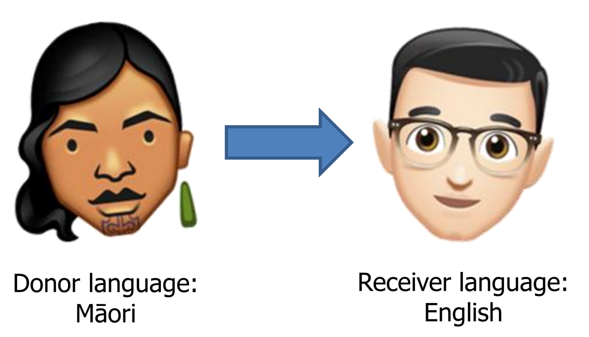
(3) I have been learning **te reo** [the Māori language] because I am a **pakeha** [European New Zealander] **roia** [lawyer] appearing in *Te Kooti Rangatahi* [youth court] and I wanted **rangatahi** [youth] many of whom are **whakama** [embarrassed] about their disconnect from their culture to recognise that this **Reo** [language] is important to us all. #LetsShareGood**TeReo**Stories (953543416352092160)

(4) We stand united Native American **Whanau** [family], **kia kaha** [be strong] #DakotaAccessPipeline #**haka** [war dance] #**Maori** #**whanau** #NativeAmerican #united (793003612217577472)

The use of Māori words has been **studied intensively** over the past thirty years, offering a comprehensive insight into the evolution of one of the youngest dialects of English – New Zealand English [1-10]. One aspect which is **missing** in this body of work is the **online discourse** presence of the loanwords – almost all studies come from (collaborative) written language (highly edited, revised and scrutinised newspaper language data [4, 8-11] and picturebooks [2-3]), or from spoken language collected in the late 1990s [7].

Loanwords are thought to arise in situations of **language contact** (i.e. when speakers of one language have contact with speakers of another). The language contact situation in New Zealand provides a unique case for loanwords, for three main reasons:

(1) The **direction of lexical transfer** is highly unusual: namely, from an endangered, indigenous language (Māori) into a dominant lingua franca (English).

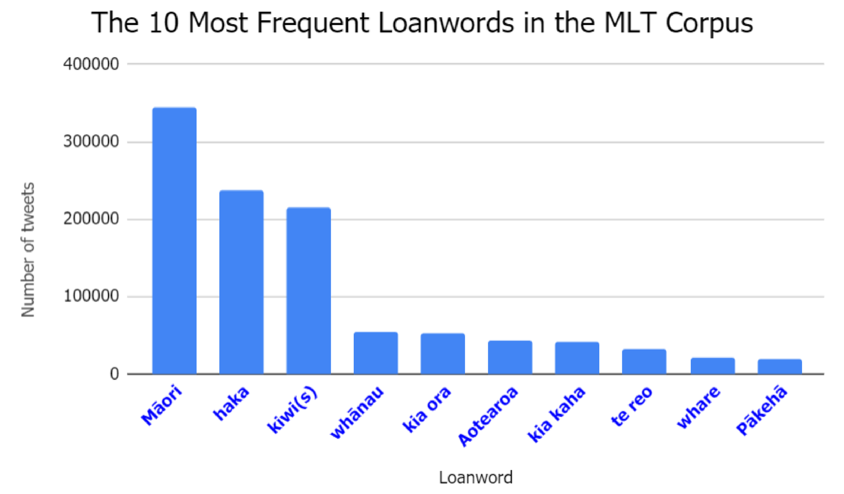


(2) Because Māori loanwords are "New Zealand's and New Zealand's alone" [12] and above speakers' consciousness, their ardent study over the years provides a fruitful comparison of the use of loanwords across **genres, contexts and time.**

(3) Loanword use is an **increasing trend** [7, 9] but the reasons for this are still unclear, and require further investigation.

The MLT Corpus

We have devised a novel method of building a corpus of New Zealand English tweets which is both (relatively) clean and large (1.2 million tweets). The *Māori Loanword Twitter Corpus* (**MLT Corpus**) affords the study of Twitter language diachronically (over an eleven-year period) and idiolectally (by user profile). To the best of our knowledge, this is the first large-scale corpus of New Zealand English tweets and the first collection of online discourse built specifically to analyse the use of Māori loanwords in New Zealand English.



1. Māori: indigenous New Zealander
2. haka: war dance
3. kiwi: New Zealander, native bird
4. whānau: family
5. kia ora: hello, thank you
6. Aotearoa: New Zealand
7. kia kaha: be strong
8. te reo: the (Māori) language
9. whare: house
10. Pakeha: European New Zealander

Problem

Our data collection method involved using target loanwords, called **query words**, to obtain (potentially) relevant tweets. After inspecting the data, it was clear that many of these query words were polysemous or otherwise unrelated to New Zealand English, and had introduced a **significant amount of noise** into the corpus. The four main types of noise are categorised below:

(1) Homographs: A word in the tweet has the **same spelling** as a loanword but a completely different meaning (e.g. **mana** is often used as a gaming term instead of the loanword meaning "pride" or "prestige").

(2) Proper nouns (with the exception of five query words that are proper nouns themselves): The loanword is used as a **personal** or **place name**, rather than a content word. Although these theoretically count as loanwords, their use does not constitute a choice (e.g. tweets containing **Moana**, meaning "sea", are dominated by references to the Disney film and princess of the same name).

(3) Misspellings: The loanword has been **mistakenly used** instead of a native English word, due to the loose and spontaneous nature of Twitter (e.g. **whare**, meaning "house", or **whero**, meaning "red", instead of English "where").

(4) Foreign languages: The tweet contains a mixture of English and **another language** that is not Māori (e.g. "mentira que voce **atua** sim! I know baby", where **atua** is a loanword meaning "God").

Our goal was to **minimise all four types of noise** in the data, so that we could test hypotheses about language change in the context of New Zealand English.

Building the Corpus

Step 1: Collect Tweets

We used the *Twitter Search API* to harvest **8 million** tweets containing one or more **query words** from a list of 116 Māori loanwords, derived from Hay [13]. The vast majority of these query words are individual words but some are short phrasal units (e.g. **kai moana**, "seafood"). We excluded most proper nouns, except those with native English counterparts (e.g. **Pākehā**, "European New Zealander").

Our search criteria are detailed below:

(1) Collect tweets posted between 2007-2018.
(2) Ensure tweets are (mostly) written in English.
(3) Convert all characters to lower-case.
(4) Remove retweets and tweets containing URLs.
(5) Remove tweets in which the query word is used as part of a username or mention (e.g. @happy_**kiwi**).
(6) For query words containing the diacritic mark for lengthened vowels, search for both the macron and non-macron variants (e.g. **māori** and **maori**).
(7) For short phrasal units, search for both the space and stripped variants (e.g. **kai moana** and **kaimoana**).
(8) Remove tweets containing fewer than five tokens (words), due to insufficient context of analysis.

The resulting collecting of tweets, termed the *Original Dataset*, was used to create the *Raw Corpus*.

Kiwi Words Website

For more information about this project, or to download the *MLT Corpus*, please visit kiwiwords.cms.waikato.ac.nz:

Step 2: Label Samples

We decided to address the "noisy" tweets in our data using supervised machine learning. Coders manually inspected a random sample of 30 tweets for each query word, and labelled each tweet as either **"relevant"** or **"irrelevant"**, depending on the loanword's context of use. Since **39 of the query words** consistently yielded irrelevant tweets (at least 90% of the time), these (and the tweets they occurred in) were **removed altogether** from the data. The annotators produced a total of **3,685 labelled tweets** for the remaining **77 query words**, which comprise the *Labelled Corpus*. Based on the assumption that the coded samples represent the real distribution of relevant/irrelevant tweets for each query word, the 39 "noisy" query words were also removed from the *Original Dataset*. In this way, we created the *Raw Corpus*, which is approximately a fifth of the size (8 million tweets reduced to 1.6 million tweets).

Step 3: Deploy Model

We **trained a classifier** using the *Labelled Corpus* as training data, so that the resulting model could be deployed on the *Raw Corpus*. Our goal was to obtain automatic predictions for the **relevance of each tweet** in this corpus, according to probabilities given by our model.

We created test and training sets that maintain the same proportion of relevant and irrelevant tweets associated with each query word in the *Labelled Corpus*. We chose to include 80% (2,949) of these tweets in the training set and 20% (736) in the test set.

Using the ***AffectiveTweets* package** [14], our labelled tweets were transformed into feature vectors based on the word n-grams they contain. We then trained various classification models on this transformed data in Weka. The models we tested were 1) **Multinomial Naive Bayes** [15] with unigram attributes and 2) L2-regularised **logistic regression** models with different word n-gram features, as implemented in LIBLINEAR [16]. We selected Multinomial Naive Bayes as the best model because it produced the highest AUC, Kappa and weighted average F-Score:

	Word *n*-grams	AUC	Kappa	F-Score
Multinomial Naive Bayes	1	**0.872**	**0.570**	**0.817**
Logistic Regression	1	0.863	0.534	0.801
	1, 2	0.868	**0.570**	0.816
	1, 2, 3	0.869	0.560	0.811
	1, 2, 3, 4	0.869	0.563	0.813
	1, 2, 3, 4, 5	0.869	0.556	0.810

Classification results on the test set.

We removed all tweets classified as irrelevant, thereby producing the *Processed Corpus*. A summary of all three corpora is given below:

	Tokens (words)	Tweets	Tweeters (authors)
Labelled Corpus	49,477	2,495	1,866
Raw Corpus	28,804,640	1,628,042	604,006
Processed Corpus	21,810,637	**1,179,390**	426,280

References

[1] Andreea Simona Calude, Steven Miller, and Mark Pagel. 2017. Modelling loanword success–a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory.*
[2] Nicola Daly. 2007. Kūkupa, koro, and kai: The use of Māori vocabulary items in New Zealand English children's picture books.
[3] Nicola Daly. 2016. Dual language picturebooks in English and Māori. *Bookbird: A Journal of International Children's Literature*, 54(3):10–17
[4] Carolyn Davies and Margaret Maclagan. 2006. Māori words–read all about it: Testing the presence of 13 Māori words in four New Zealand newspapers from 1997 to 2004. *Te Reo*, 49.
[5] Julia De Bres. 2006. Maori lexical items in the mainstream television news in New Zealand. *New Zealand English Journal*, 20:17.
[6] Marta Degani and Alexander Onysko. 2010. Hybrid compounding in New Zealand English. *World Englishes*, 29(2):209–233.
[7] Graeme Kennedy and Shunji Yamazaki. 1999. The influence of Maori on the New Zealand English lexicon. *Language and Computers*, 30:33–44.
[8] John Macalister. 2009. Investigating the changing use of Te Reo. *NZ Words*, 13:3–4.
[9] John Macalister. 2006a. The Maori lexical presence in New Zealand English: Constructing a corpus for diachronic change. Corpora, 1(1):85–98.
[10] Alexander Onysko and Andreea Calude. 2013. Comparing the usage of Māori loans in spoken and written New Zealand English: A case study of Māori, Pākehā, and Kiwi. *New perspectives on lexical borrowing: Onomasiological, methodological, and phraseological innovations*, pages 143–170.
[11] John Macalister. 2006b. the Maori presence in the New Zealand English lexicon, 1850–2000: Evidence from a corpus-based study. *English WorldWide*, 27(1):1–24.
[12] Tony Deverson. 1991. New Zealand English lexis: the Maori dimension. *English Today*, 7(2):18–25.
[13] Jennifer Hay. 2018. What does it mean to "know a word?". In Language and Society Conference of NZ in November 2018 in Wellington, NZ.
[14] Felipe Bravo-Marquez, Eibe Frank, Bernhard Pfahringer, and Saif M. Mohammad. 2019. AffectiveTweets: a Weka package for analyzing affect in tweets. *Journal of Machine Learning Research*, 20:1–6.
[15] Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.
[16] https://www.csie.ntu.edu.tw/~cjlin/liblinear/