What can Social Media tell us about Māori Loanwords?

David Trye & Andreea Calude
(with Felipe Bravo-Márquez and Te Taka Keegan)

Background

...where we are starting from

Two main waves of borrowing, we are possibly still in the peak of the second wave, or at the start of another wave.

Loanword use – both in term and speaker/writer

Some loanwords ‘do’ better than others: shorter than native English counterparts, core (rather than cultural).


Loanword use is skewed, by topic and speaker/writer (ethnicity and gender).

Loanword use is increasing – in those texts that use them – both in terms of types and tokens.

Some Māori loanwords projects

MATARIKI CORPUS (with Sally Harper, Steven Miller & Hēmi Whanaga)
2007-2016
~ 91,958 words
~ 194 articles
→ 2,673 loanword tokens / 282 loanword types, rate 29/1,000 words

MĀORI LANGUAGE WEEK CORPUS (with Katie Levendis)
2008-2017
~ 108,925 words
~ 290 articles
→ 3,795 loanword tokens / 186 loanword types, rate 35/1,000 words

NATIONAL SCIENCE CHALLENGE CORPUS (with Louise Stevenson, Hēmi Whanaga & Te Taka Keegan)
Snapshot in Jan 2018
~ 1.5 million words
~ 12 websites & 11 Twitter feeds
→ ?? loanword tokens / ?? loanword types, rate ??/1,000 words

MĀORI LOANWORDS TWITTER CORPUS (with David Trye, Felipe Bravo & Te Taka Keegan)
A lot of what we know comes from newspaper data:
- **FORMAL**
- **HIGHLY EDITED**
- **PRESCRIPTIVE**
- **COLLABORATIVE**
- **NORMATIVE**

**CHEAP TO GET LOTS OF IT**
- **FORMAL & INFORMAL**
- **NOT EDITED**
- **CREATIVE**
- **SINGLE-AUTHORED**
- **NORMATIVE & NON-NORM**

---

**Twitter**

A lot of what we know comes from newspaper data:
- **FORMAL**
- **HIGHLY EDITED**
- **PRESCRIPTIVE**
- **COLLABORATIVE**
- **NORMATIVE**

...where we detail corpus building strategy and rationale

Overview

The Māori Loanword Twitter Corpus (MLT Corpus) is a collection of NZE tweets containing Māori loanwords.

We devised a new method for building a corpus that is sufficiently large, clean and balanced, consisting of three main steps:

1. Collect Tweets
   - Proud to be a **kiwi**
   - Love my crazy **whanau**
   - Moana is my fav Princess
   - haka ne kuma fa

2. Annotate Tweets
   - Proud to be a kiwi
   - Love my crazy whanau
   - Moana is my fav Princess
   - haka ne kuma fa

3. Deploy Model
   - tweet
   - tweet
   - tweet
   - tweet
Building the **MLT Corpus** (I)

116 Query Words

- Aotearoa
- Aroha
- Atua
- Awa
- ...
- Whero

8 Million Tweets

**Twitter Data**

<table>
<thead>
<tr>
<th>tweet_id</th>
<th>username</th>
<th>timestamp (GMT)</th>
<th>query word</th>
<th>text</th>
</tr>
</thead>
<tbody>
<tr>
<td>7573693436</td>
<td>JustStephOK</td>
<td>2016-07-25 12:18</td>
<td>waiata</td>
<td>Led the <strong>waiata</strong> for the manuhiri at the pōwhiri for new staff for induction week. Was told by the kaumātua I did it with mana &amp; integrity.</td>
</tr>
</tbody>
</table>
Building the MLT Corpus (I)

116 Query Words
Aotearoa
Aroha
Atua
Awa

8 million Tweets

4,600 Labelled Tweets
Proud to be a kiwi
Love my crazy whanau
Moana is my fav Princess

Common Types of Noise

1) Homographs
   • Same spelling as loanword but different meaning
   • e.g. mana used as gaming term instead of pride/prestige

2) Proper Nouns
   • Personal/place name used (rather than content word)
   • Theoretically count as loanwords, but their use does not constitute a choice
   • e.g. Moana used to refer to Disney princess/film

3) Misspellings
   • Loanword mistakenly used instead of native English word
   • Result of impromptu/spontaneous nature of Twitter
   • e.g. whare or whero instead of “where”

4) Foreign Languages
   • Tweet contains English and some other language that is not Māori
   • e.g. “mentira que voce atua sim! I know baby”
Building the **MLT Corpus** (i)

116 Query Words
- Aotearoa
- Aroha
- Atua
- Awa
- Whero

8 million Tweets

<table>
<thead>
<tr>
<th>Raw Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>77 'best' loanwords</td>
</tr>
<tr>
<td>1.6 million Tweets</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Labelled Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>4,600 Labelled Tweets</td>
</tr>
<tr>
<td>77 'best' loanwords</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Processed Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2 million Tweets</td>
</tr>
</tbody>
</table>

Building the **MLT Corpus** (ii)

<table>
<thead>
<tr>
<th>Vocab</th>
</tr>
</thead>
<tbody>
<tr>
<td>word, word, ... word,</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tweet Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>word, word, ... word,</td>
</tr>
<tr>
<td>tweet_1 0 1 ... 0</td>
</tr>
<tr>
<td>tweet_2 1 0 ... 1</td>
</tr>
<tr>
<td>... ... ... ...</td>
</tr>
<tr>
<td>tweet_m 1 1 ... 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tweet Vectors</th>
</tr>
</thead>
<tbody>
<tr>
<td>tweet, tweet, ... tweet,</td>
</tr>
<tr>
<td>tweet_1 OK</td>
</tr>
<tr>
<td>tweet_2 X</td>
</tr>
<tr>
<td>tweet_3 OK</td>
</tr>
<tr>
<td>...</td>
</tr>
<tr>
<td>tweet_m OK</td>
</tr>
</tbody>
</table>
**Machine Learning**

**Input and Output**

### Training Data (Input)

<table>
<thead>
<tr>
<th>id</th>
<th>username</th>
<th>timestamp</th>
<th>query word</th>
<th>text</th>
<th>relevance label</th>
</tr>
</thead>
<tbody>
<tr>
<td>7573693</td>
<td>JustStephOK</td>
<td>2016-07-25 12:18</td>
<td>waiata</td>
<td>Led the <strong>waiata</strong> for the manuhiri at the pōwhiri for new staff for induction week. Was told by the kaumātua I did it with mana &amp; integrity.</td>
<td>Relevant</td>
</tr>
</tbody>
</table>

### Target Data (Output)

<table>
<thead>
<tr>
<th>id</th>
<th>username</th>
<th>timestamp</th>
<th>query word</th>
<th>text</th>
<th>prob_rel</th>
</tr>
</thead>
<tbody>
<tr>
<td>8095892</td>
<td>KUOI_DJ</td>
<td>2016-12-16 15:41</td>
<td>waiata</td>
<td>Split Enz—History Never Repeats— <strong>Waiata</strong></td>
<td>0.078 (irrelevant)</td>
</tr>
</tbody>
</table>

### Classification Results

<table>
<thead>
<tr>
<th></th>
<th>Word n-grams</th>
<th>AUC</th>
<th>Kappa</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multinomial Naive Bayes</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.872</td>
<td>0.570</td>
<td>0.817</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.863</td>
<td>0.534</td>
<td>0.801</td>
<td></td>
</tr>
<tr>
<td>1, 2</td>
<td>0.868</td>
<td>0.570</td>
<td>0.816</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>0.869</td>
<td>0.560</td>
<td>0.811</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>0.869</td>
<td>0.563</td>
<td>0.813</td>
<td></td>
</tr>
<tr>
<td>1, 2, 3, 4, 5</td>
<td>0.869</td>
<td>0.556</td>
<td>0.810</td>
<td></td>
</tr>
</tbody>
</table>

| **Logistic Regression** |              |        |        |           |
| 1                    | 0.872        | 0.570  | 0.817  |
## Corpus Statistics

<table>
<thead>
<tr>
<th></th>
<th>Tokens (words)</th>
<th>Tweets</th>
<th>Tweeters (authors)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Labelled Corpus</strong></td>
<td>49,477</td>
<td>2,495</td>
<td>1,866</td>
</tr>
<tr>
<td><strong>Raw Corpus</strong></td>
<td>28,804,640</td>
<td>1,628,042</td>
<td>604,006</td>
</tr>
<tr>
<td><strong>Processed Corpus</strong></td>
<td>21,810,637</td>
<td>1,179,390</td>
<td>426,280</td>
</tr>
</tbody>
</table>

What we are doing now...  
... where we discuss current analyses
Hybrid hashtags

First used in August 2007, by Chris Messina – “how do you feel about using # (pound) for groups. As in #barcamp [msg]?”

(cf. Caleffi 2015)

"a way to categorize messages posted on Twitter" (Cunha et al 2011:58)

"paralanguage of Twitter" (Maity, Saraf and Mukherjee, 2016:1)

"community building linguistic activity" (Zappavigna, 2011:2)

"enable ambient affiliation" (Zappavigna, 2012: 790)

"a crucial currency which enables visibility and projects potential interaction with members of the site" (Page, 2012:6)

"referring to a topic and creating communities of people interested in [a] topic" (Caleffi 2015:67)
Hashtags – how do they arise?

Caleffi (2015) – new morphological process, not a compound, not blending, not agglutination – *hashtags*.

Maity, Saraf and Mukkerjee (2016) – compounding, more common than in “standard texts and language”.

Hashtags – ingredients of popular #s?

‘SUCCESSFUL’ hashtags

(Cunha et al 2011)

Simple, Direct

Short

No underscore

‘SUCCESSFUL’ compound hashtags (Maity, Saraf and Mukkerjee 2016)

PropN–PropN > comN–comN > det-comN > V-det

Listed-listed > unlisted-unlisted > listed-unlisted

Word overlap (check the overlap in the words in tweets with #A vs. tweets with #B)

n-gram overlap (segment hashtag A & B, find tweets with #A and #B, check for 2-grams, 3-grams, for each all words in these tweets and d then check for overlapping n-grams for tweets containing #A and tweets containing #B)

(Total number of words in the hashtag was not as important)

TOPICS: Association with hashtags with controversial politics and sport makes them more sticky and more pervasive.
Hybrid hashtags (n=38) labelled corpus (45K words)

Hybrid hashtags (n=60) processed corpus freq>15
Hybrid hashtags (n=10) processed corpus freq>150

Hybrid hashtags (n=60) processed corpus freq>15
Hybrid hashtags (n=60) processed corpus freq>15

$\text{Number of Tweets} = 0.91, p < 2.2e-16$

Whakapapa

[Graph showing correlation between number of distinct users and number of tweets, with $r = 0.91, p < 2.2e-16$]
Acknowledgements

Tēnā koutou katoa!