#KiwiIngenuity – Creative Uses of Māori Loanwords in NZE Twitter Posts

David Trye, Andreea Calude, Felipe Bravo Marquez & Te Taka Keegan

University of Waikato
University of Chile

What can social media tell us about our use of Māori loanwords in NZE?

Kia kaha All Blacks hard. Loss! England well done keeping the the old steam train running on full from the start and the old v stance I’m gud with it being a kiwi Maori #worldinunion #ENGvNZL

8:27 AM · Oct 27, 2019 · Twitter for iPhone
Loanwords in NZE

Many genres studied (De Bres, Daly, Macalister, Davies & Maclagan, Degani, Onysko).

For good reason: unusually productive lexical transfer situation from minority language to dominant language.

Two main waves of borrowing (Macalister 2006).

Loanword use remains an increasing trend today, in both types and tokens (especially social culture terms).

Use is skewed across speakers (Māori females lead the change) & across topics (Māori-related topics draw highest counts).

Rationale – Why Twitter?

Newspaper Data
- Formal
- Highly Edited
- Prescriptive
- Collaborative
- Normative

Twitter Data
- Formal & Informal
- Not Edited
- Creative
- Single-Author
- Normative & Non-normative

PLUS
- Cheap to get
- Lots of it

BUT NOISY !!!!
Building the MLT Corpus (1)

116 loanwords
“query words”
Aotearoa
Aroha
Atua
Awa
...
Whereo

8 million tweets
(2007-2018)

4,600 labelled tweets
40 tweets per query word

Proud to be a kiwi
Love my crazy whanau
Moana is my fav Princess
haka ne kuma fa you say

Raw Corpus
77 ‘best’ loanwords
~4.5 million tweets

Labelled Corpus
77 ‘best’ loanwords
3,600 labelled tweets

Building the MLT Corpus (2)

Labelled Corpus

Tweet Vectors
word_1 word_2 ... word_n
tweet_1 0 1 ... 0
tweet_2 1 0 ... 1
... ... ... ...
tweet_m 1 1 ... 0

MLT Corpus
~2.8 million tweets

Raw Corpus

tweet_1 ✓
tweet_2 ✗
tweet_3 ✓
... ✓
tweet_M ✓
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>75736934366442480640</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>809589244037566460</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</td>
</tr>
</tbody>
</table>

Classification Results

<table>
<thead>
<tr>
<th>Word n-grams</th>
<th>Linear Logistic Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AUC</td>
</tr>
<tr>
<td>1</td>
<td>0.863</td>
</tr>
<tr>
<td>1, 2</td>
<td><strong>0.868</strong></td>
</tr>
<tr>
<td>1, 2, 3</td>
<td>0.869</td>
</tr>
<tr>
<td>1, 2, 3, 4</td>
<td>0.869</td>
</tr>
<tr>
<td>1, 2, 3, 4, 5</td>
<td>0.869</td>
</tr>
</tbody>
</table>

@attribute WNGRAM-1-indifferent numeric
@attribute WNGRAM-2-coming-up numeric

...
Corpus Overview

<table>
<thead>
<tr>
<th></th>
<th>Tokens (words)</th>
<th>Tweets</th>
<th>Tweeters (authors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labelled Corpus (Rel.)</td>
<td>~50,000</td>
<td>~2,500</td>
<td>~1,900</td>
</tr>
</tbody>
</table>

PLUS statistically significant, increasing, diachronic trends for 55 of the 77 query words, suggesting that loanword use on Twitter is increasing for most tokens.

Hybrid Hashtags

Hashtags containing lexical items from both English and Māori.
Syntactic Word Class

ADJP = adjective phrase (n=3)
ADVP = adverb phrase (n=2)
CLAUSE = NP + NP (n=6)
CNP = common NP (n=42)
FORMULAIC = formula (n=5)
PNP = proper NP (n=7)
VP = verb phrase (n=13)
NA = unsure (n=3)

Flora & fauna (n=5)
Generic (n=3)
humour (n=4)
Maori culture (n=18)
NZ identity (n=35)
Sport (n=15)
NA = unsure (n=1)
Word Length

Diachronic Trends (Year)
Sum of area = 592
50% of that = 296
Half-life = 4 years

Comparing Half-life and Persistence

Persistence (num of months it is used)

Half-life (num months it takes to cover half its activity)

Background  MLT Corpus (Data)  Hybrid Hashtags (Findings)  Where to next?
Predicting Half-life (persistence)

Background     ML T Corpus (Data)     Hybrid Hashtags (Findings)     Where to next?

Findings & Future Directions

Twitter is a valuable source of language data for studying loanwords in informal, unedited discourse.

Machine learning techniques and the use of query words can be fruitfully used to create mixed-language corpora.

One of the most distinctive uses of loanwords in these data appears to be within hybrid hashtags.

These hybrid hashtags are linguistically diverse:
- They are different lengths (2-6 words)
- Their word class varies (nouns, adj, adv, verb, clauses)
- They pertain to different semantic domains (including NZ identity, Māori culture, sport & humour).

Hybrid hashtags are comprised of many English word-types but only a handful of loanword-types.

From the 77 query words searched → 81 hybrid hashtags → 9 loans

Partaking in hybrid hashtags may be another predictor of entrenchment.

» Position of the hybrid hashtag within the tweet
» Number of followers for each user
Word Embeddings

Distributional Hypothesis:
“You shall know a word by the company it keeps”

Word2Vec

Similar words receive similar vectors, and are situated closer together in n-dimensional space.
### Recent Corpora

**Matariki Corpus**  
(with Sally Harper, Steven Miller & Hēmi Whaanga)  
» 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 Whaanga & Te Taka Keegan)  
» Snapshot in Jan 2018  
» ~ 1.5 million words  
» ~ 12 websites & 11 Twitter feeds

**Māori Loanword Twitter (MLT) Corpus**  
with David Trye, Felipe Bravo Marquez & Te Taka Keegan

### Background

ML T Corpus (Data)  
Hybrid Hashtags (Findings)  
Where to next?

### Comparisons of Loanword Use with Other Corpora (Top 10)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Maori</td>
<td>Māori</td>
<td>Māori</td>
<td>Kiwi(s)</td>
<td>Kiwi(s)</td>
</tr>
<tr>
<td>Pakeha</td>
<td>Māori</td>
<td>te reo</td>
<td>Māori</td>
<td>Māori</td>
</tr>
<tr>
<td>Kiwi</td>
<td>marae</td>
<td>iwi</td>
<td>haka</td>
<td>haka</td>
</tr>
<tr>
<td>marae</td>
<td>puanga</td>
<td>reo</td>
<td>kia ora</td>
<td>(te) reo</td>
</tr>
<tr>
<td>kakapo</td>
<td>kapa haka</td>
<td>whānau</td>
<td>Whānau</td>
<td>hui</td>
</tr>
<tr>
<td>tangata</td>
<td>te reo</td>
<td>marae</td>
<td>Aotearoa</td>
<td>Waitangi</td>
</tr>
<tr>
<td>tiriti</td>
<td>whānau</td>
<td>maka haka</td>
<td>kia kaha</td>
<td>Aotearoa</td>
</tr>
<tr>
<td>pa</td>
<td>waka</td>
<td>Pakeha</td>
<td>môrena</td>
<td>kai</td>
</tr>
<tr>
<td>whenua</td>
<td>hangi</td>
<td>Kiwi</td>
<td>te reo</td>
<td>kia ora</td>
</tr>
<tr>
<td>aroha</td>
<td>iwi</td>
<td>kia ora</td>
<td>whare</td>
<td></td>
</tr>
</tbody>
</table>
Which Hashtags “Survive” Longer?

Using Half-lifes

Length of hashtag (number of words)  Semantic Domain  Word Class
Half-life = 4 years

Sum of area =
0+7+17+74+125+109+61+7+1+68+60 = 592

50% of area = 592/2 = 296

0+7+17+74+125 =

Half-life = 4 years