Collotextualisation
An alternative approach to studying loanwords

David Trye & Andreea S. Calude
University of Waikato

Rationale

• Loanword use is typically investigated using frequency-based measures, such as number of types and tokens in a corpus
  • However, it makes sense to also study groups of loanwords, not just loanwords in isolation

• Anecdotal evidence suggests that New Zealand English (NZE) texts either contain several Māori loanwords or none at all
  • As observed in children’s picture books (Macdonald & Daly, 2013)

• We need a way to capture these groupings, including loanwords that are dispersed throughout the text (not necessarily in close proximity)
Concordances

Source: AntConc (Anthony, 2020)

Key Terms

Collocation (Firth, 1957)
- The company a word keeps
- Collocation networks produced by tools like #LancsBox (Brezina, 2020)

Collostruction (Stefanowitsch & Gries, 2003)
- Words that co-occur with certain grammatical constructions
- Captures interactions between the lexicon and syntax

Collotextualisation (this talk)
- Words that co-occur anywhere within the same text, regardless of length
- Captures interactions between the lexicon and the greater discourse
Māori Loanwords in New Zealand English

- Many genres have been studied
- Unusually productive lexical transfer situation from minority language to dominant language
- Two main waves of borrowing (Macalister, 2006)
- Loanword use is increasing
  - With respect to both types and tokens
  - Especially social culture terms
- Use skewed across speakers and topics
  - Māori females lead the change
  - Māori-related topics draw highest counts

What can collotextualisation tell us about Māori loanword use in NZE?
Data Summary

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Citation</th>
<th>Tokens</th>
<th>Texts</th>
<th>Average Tokens per Text</th>
<th>Loanwords per 1,000 Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matariki Corpus (2007-2016)</td>
<td>Calude et al., 2019</td>
<td>91,958</td>
<td>194</td>
<td>474</td>
<td>29</td>
</tr>
<tr>
<td>Press Corpus (1996-2011)</td>
<td>Calude &amp; James, 2011</td>
<td>5.1 million</td>
<td>990</td>
<td>5,158.6</td>
<td>5</td>
</tr>
</tbody>
</table>

Extracting Māori N-grams

- Using code available online, we generated a list of valid Māori words present in each corpus
  - [https://github.com/TeHikuMedia/ nga-kupu](https://github.com/TeHikuMedia/ nga-kupu)

- The output was modified to also include multi-word n-grams (e.g. te reo Māori, kapa haka, kia ora e te whānau)

- We then sorted these words & phrases by frequency
  - N-grams embedded inside other n-grams were only counted once (as part of the largest n-gram)
Cleaning the Data

- 140 loanwords/phrases after discarding irrelevant & infrequent n-grams
- N-grams were coded for the following linguistic properties:
  1. Semantic Category
  2. Semantic Type
  3. Size/length (words)
  4. Listedness

Computing N-gram Sets

<table>
<thead>
<tr>
<th>Texts</th>
<th>Loanwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>1 0 ... L140</td>
</tr>
<tr>
<td>T2</td>
<td>0 1 ... 0</td>
</tr>
<tr>
<td>...</td>
<td>... ... ...</td>
</tr>
<tr>
<td>TN</td>
<td>0 1 ... 1</td>
</tr>
</tbody>
</table>

Background Data & Methods Findings Future Work
Network Metrics

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>n-gram</td>
<td>total_sets</td>
<td>distinct_sets</td>
<td>degree</td>
<td>avg_freq</td>
<td>avg_length</td>
<td>trigger_occurrences</td>
<td>trigger_ratio</td>
</tr>
<tr>
<td>2</td>
<td>maori</td>
<td>274</td>
<td>229</td>
<td>102</td>
<td>1.19990959</td>
<td>6.20524017</td>
<td>97</td>
<td>0.394014985</td>
</tr>
<tr>
<td>3</td>
<td>te reo</td>
<td>165</td>
<td>154</td>
<td>96</td>
<td>1.261268701</td>
<td>6.64287743</td>
<td>86</td>
<td>0.484048694</td>
</tr>
<tr>
<td>4</td>
<td>te reo maori</td>
<td>175</td>
<td>146</td>
<td>87</td>
<td>1.166930137</td>
<td>6.9692197</td>
<td>82</td>
<td>0.1028571429</td>
</tr>
<tr>
<td>5</td>
<td>te whai o te reo maori</td>
<td>64</td>
<td>89</td>
<td>83</td>
<td>1.96382256</td>
<td>7.16275907</td>
<td>16</td>
<td>0.370212766</td>
</tr>
<tr>
<td>6</td>
<td>tiki</td>
<td>45</td>
<td>82</td>
<td>56</td>
<td>1.047659481</td>
<td>7.880521861</td>
<td>7</td>
<td>0.135536354</td>
</tr>
<tr>
<td>7</td>
<td>aotearoa</td>
<td>44</td>
<td>38</td>
<td>61</td>
<td>1.157894737</td>
<td>6.88421653</td>
<td>3</td>
<td>0.6818181818</td>
</tr>
<tr>
<td>8</td>
<td>reo</td>
<td>30</td>
<td>30</td>
<td>60</td>
<td>1</td>
<td>0.933333333</td>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>9</td>
<td>whanau</td>
<td>39</td>
<td>39</td>
<td>65</td>
<td>1</td>
<td>8.333333333</td>
<td>3</td>
<td>0.65128205128</td>
</tr>
<tr>
<td>10</td>
<td>te taura whiri i te reo</td>
<td>30</td>
<td>34</td>
<td>52</td>
<td>1.059555556</td>
<td>6.888888889</td>
<td>1</td>
<td>0.2631578947</td>
</tr>
<tr>
<td>11</td>
<td>mara</td>
<td>35</td>
<td>35</td>
<td>60</td>
<td>1</td>
<td>8.25742857</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>kapa haka</td>
<td>25</td>
<td>23</td>
<td>59</td>
<td>1</td>
<td>0.36</td>
<td>3</td>
<td>0.12</td>
</tr>
<tr>
<td>13</td>
<td>pakeha</td>
<td>20</td>
<td>25</td>
<td>49</td>
<td>1</td>
<td>0.74</td>
<td>1</td>
<td>0.0346153846</td>
</tr>
<tr>
<td>14</td>
<td>kohanga reo</td>
<td>29</td>
<td>20</td>
<td>56</td>
<td>1</td>
<td>0.286714286</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>kura</td>
<td>15</td>
<td>15</td>
<td>60</td>
<td>1</td>
<td>0.933333333</td>
<td>1</td>
<td>0.0696666667</td>
</tr>
<tr>
<td>16</td>
<td>kuki</td>
<td>15</td>
<td>15</td>
<td>28</td>
<td>1</td>
<td>5.486666667</td>
<td>5</td>
<td>0.333333333</td>
</tr>
<tr>
<td>17</td>
<td>manaakitanga</td>
<td>6</td>
<td>6</td>
<td>14</td>
<td>1</td>
<td>6.5</td>
<td>2</td>
<td>0.333333333</td>
</tr>
<tr>
<td>18</td>
<td>non-maori</td>
<td>23</td>
<td>21</td>
<td>40</td>
<td>1.065230095</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>19</td>
<td>mataniki</td>
<td>13</td>
<td>12</td>
<td>25</td>
<td>1.083333333</td>
<td>5.833333333</td>
<td>3</td>
<td>0.2307692308</td>
</tr>
<tr>
<td>20</td>
<td>haka</td>
<td>16</td>
<td>16</td>
<td>32</td>
<td>1</td>
<td>6.3125</td>
<td>1</td>
<td>0.0625</td>
</tr>
</tbody>
</table>

N-grams per Text

Our corpora have many more texts with two or more loanwords than only one
Visualising Pairwise Relationships

- Co-occurrence network
- Nodes = Loanwords
- Node Size = Frequency
- Links / Edges = Text-level co-occurrence
- Colour = Semantic type
  (orange: core, purple: cultural, green: N/A)

Visualising Sets

- On average, each n-gram occurs in a set with 6 others
- Infrequent loans cluster around frequent loans
  - If a text contains “mokopuna”, it will also likely contain “Māori” and “te reo”
  - This clustering suggests that loanwords might occur in vocabulary frequency bands
    (as proposed for measuring L2 vocabulary; see Laufer & Nation, 1995)
Total Sets vs N-gram Frequency

N-gram frequency is positively correlated with set frequency (frequent loanwords occur in more sets)

Trigger Loan (TL) Analysis

- Frequent n-grams are typically the first to occur in a given text
- Māori, Kiwi, Matariki and te reo constitute 60% of all trigger loan occurrences
• 54% of trigger loan occurrences (471/870) are PNs of length 1
  - Māori, Pākehā, Matariki, Kiwi, Aotearoa, Aoraki
• Most trigger loan tokens are PN (60%) or SC (28%)
• Shorter n-grams more common (83% = 1 word, 11% = 2 words)
• 21% core, 19% cultural
• 80% listed in dictionary

Currently...

• Cluster analysis
• Set characteristics
• Diachronic patterns
• Corpus / genre comparison

• In future:
  • Are there other groups of words that might benefit from a macro-discourse approach?
Wrapping Up

• We propose a new method for studying loanwords at a macro-discourse level, called *collotextualisation*
  • Instead of limiting the analysis to nearby collocates, we extract all loanwords that co-occur within the same text
  • Each text therefore contains a set of loanwords; these sets can be analysed quantitatively
• When it comes to Māori loanwords in NZE...
  • The use of one loanword is likely to trigger the use of others (often several) in the same text
  • Speakers do not appear to make individual word choices but rather adopt loanwords as a set (motivated by ideology?)

Thank you!

Special thanks to:

• Te Hiku Media
• Sally Harper
• Katie Levendis
• The University of Waikato for funding this research through a Doctoral Scholarship