

Collotextualisation

An alternative approach to studying loanwords

David Trye & Andreea S. Calude
University of Waikato

1

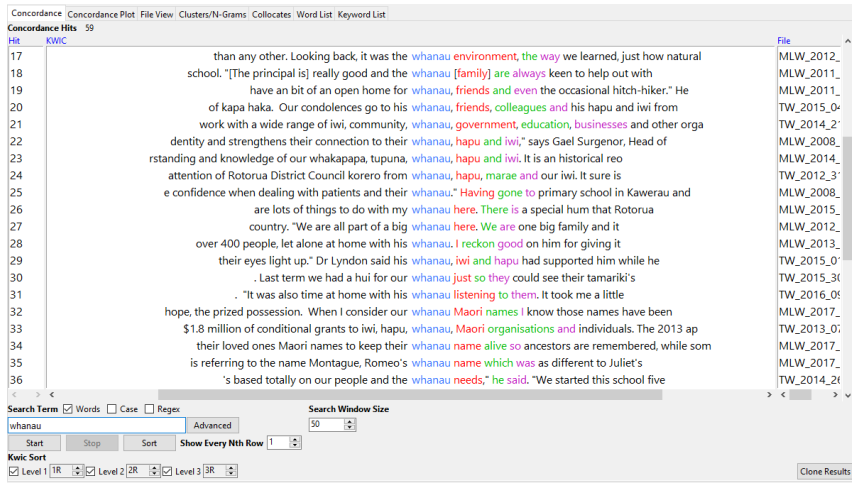
Rationale

- 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)

Background Data & Methods Findings Future Work

2

Concordances



Source: AntConc (Anthony, 2020)

Background Data & Methods Findings Future Work

3

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*

Background Data & Methods Findings Future Work

4

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



Kiwi



Whānau

Visual Evolution
illustration + design

Background Data & Methods Findings Future Work




5

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

Background Data & Methods Findings Future Work

6

Data Summary

Corpus	Citation	Tokens	Texts	Average Tokens per Text	Loanwords per 1,000 Tokens
 MLW Corpus (2008-2017)	Levendis & Calude, 2019	108,521	289	375.5	35
 Matariki Corpus (2007-2016)	Calude et al., 2019	91,958	194	474	29
 Press Corpus (1996-2011)	Calude & James, 2011	5.1 million	990	5,158.6	5

Background [Data & Methods](#) Findings [Future Work](#)

7

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>
- 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)



Background [Data & Methods](#) Findings [Future Work](#)

8

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



Background [Data & Methods](#) Findings Future Work

9

Computing N-gram Sets

Loanwords

		Loanwords			
		L1	L2	...	L140
Texts	T1	1	0	...	1
	T2	0	1	...	0

	TN	0	1	...	1

Background [Data & Methods](#) Findings Future Work

10

Network Metrics

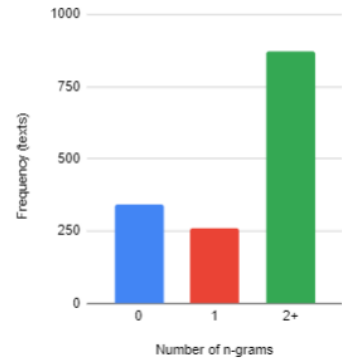
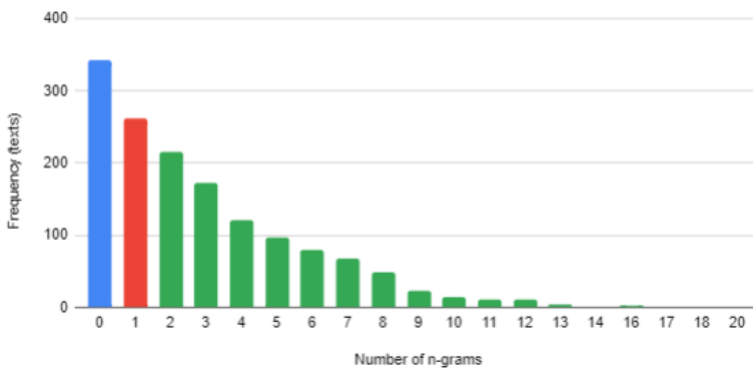
	A	J	K	L	M	N	O	P
1	n-gram	total_sets	distinct_sets	degree	avg_freq	avg_length	trigger_occurrences	trigger_ratio
2	maori	274	229	102	1.19650655	6.205240175	97	0.3540145985
3	te reo	185	154	96	1.201298701	6.642857143	86	0.4648648649
4	te reo maori	175	146	87	1.198630137	6.808219178	32	0.1828571429
5	te wiki o te reo maori	94	86	83	1.093023256	7.11627907	16	0.170212766
6	iwi	44	42	56	1.047619048	7.880952381	6	0.1363636364
7	aotearoa	44	38	61	1.157894737	6.868421053	3	0.06818181818
8	reo	30	30	57	1	8.933333333	3	0.1
9	whanau	39	39	65	1	8.333333333	2	0.05128205128
10	te taura whiri i te reo	38	36	52	1.055555556	6.888888889	1	0.02631578947
11	marae	35	35	60	1	8.257142857	0	0
12	kapa haka	25	25	56	1	8.36	3	0.12
13	pakeha	26	25	49	1.04	7.4	1	0.03846153846
14	kohanga reo	28	28	56	1	9.285714286	0	0
15	kura	15	15	39	1	8.933333333	1	0.06666666667
16	kiwi	15	15	28	1	5.466666667	5	0.3333333333
17	manaakitanga	6	6	14	1	6.5	2	0.3333333333
18	non-maori	23	21	40	1.095238095	8	0	0
19	matariki	13	12	25	1.083333333	5.833333333	3	0.2307692308
20	haka	16	16	32	1	6.3125	1	0.0625

Background [Data & Methods](#) Findings [Future Work](#)

11

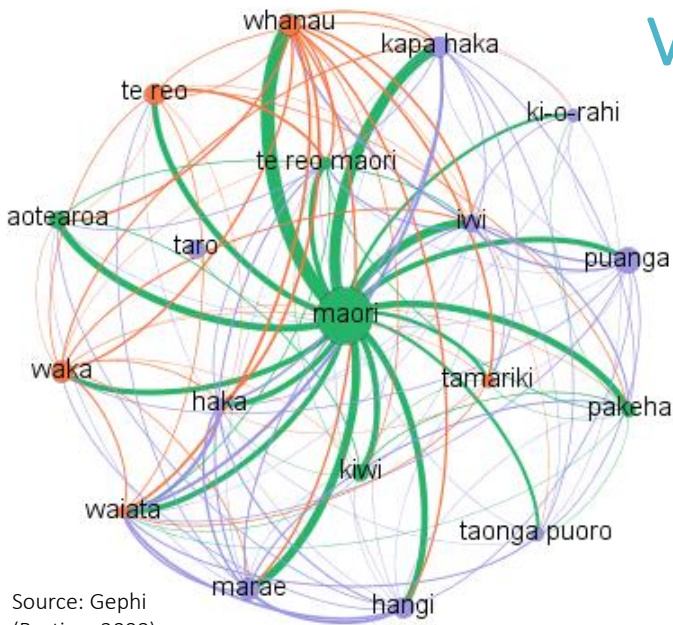
N-grams per Text

Our corpora have many more texts with *two or more* loanwords than only one



Background [Data & Methods](#) [Findings](#) [Future Work](#)

12



Source: Gephi
(Bastian, 2008)

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)

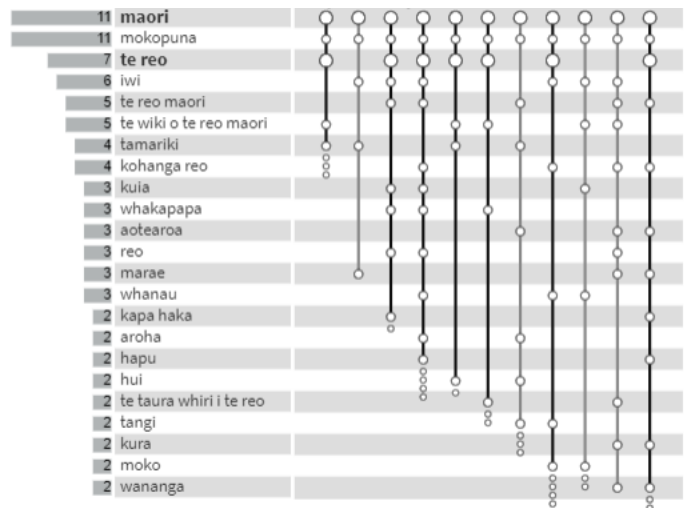
Background Data & Methods Findings Future Work

13

Visualising Sets

Source: PaohVis
(Valdivia et al., 2019)

- 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)



Background Data & Methods Findings Future Work

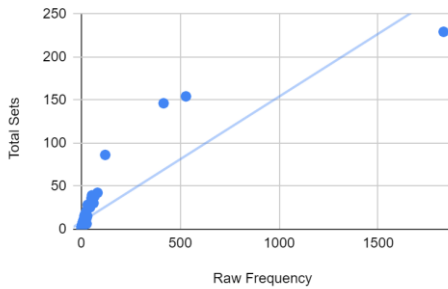
14

Total Sets vs N-gram Frequency

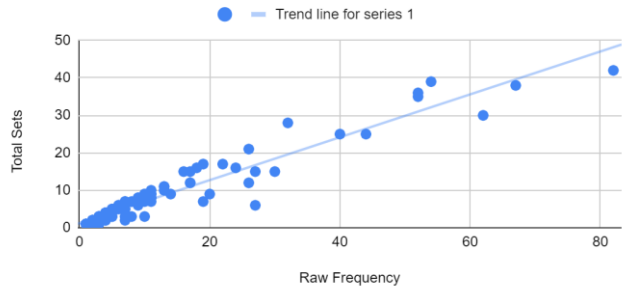


N-gram frequency is positively correlated with set frequency (frequent loanwords occur in more sets)

Total Sets vs Raw Frequency



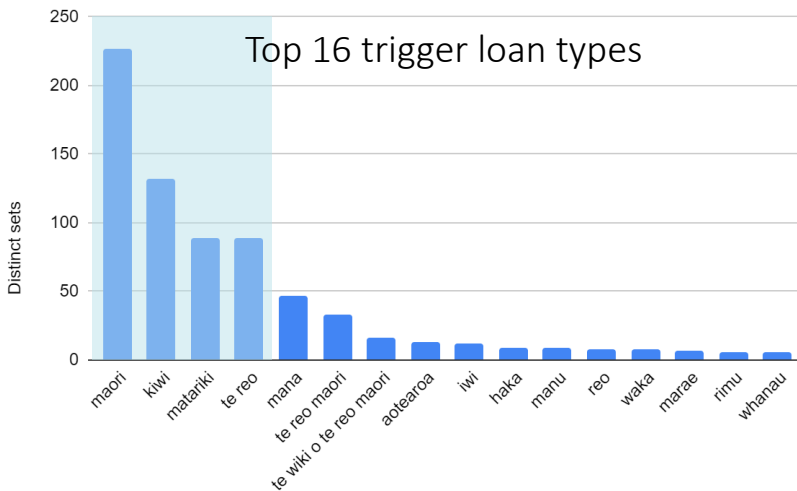
Total Sets vs Raw Frequency



Background Data & Methods [Findings](#) Future Work

15

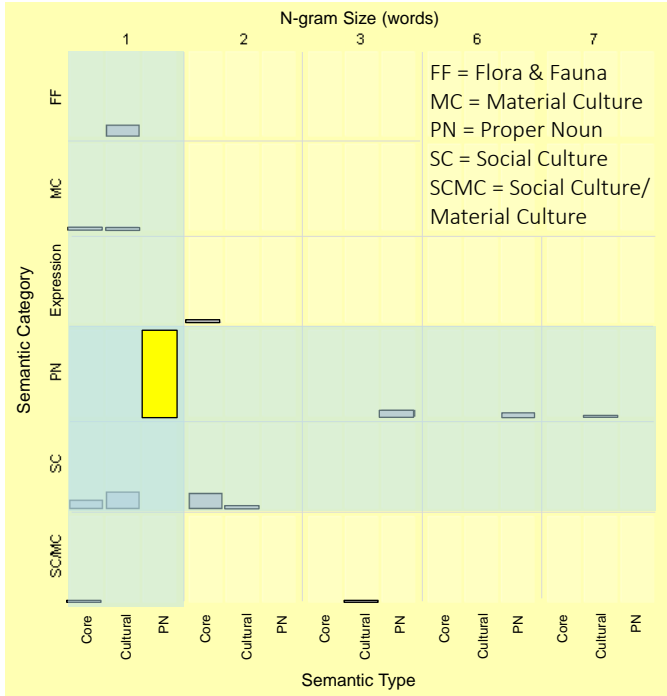
Trigger Loan (TL) Analysis



- 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

Background Data & Methods [Findings](#) Future Work

16



Source: Mondrian
(Theus, 2010)

TL Tokens

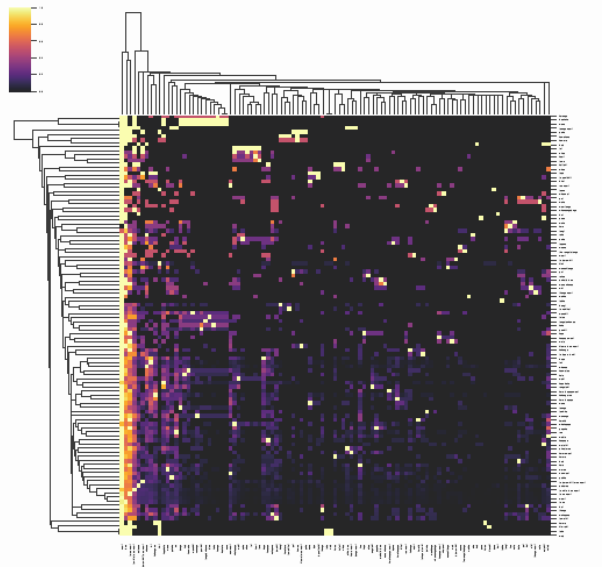
- 54% of **trigger loan occurrences** (471/870) are PNs of length 1
 - Māori, Pākehā, Matariki, Kiwi, Aotearoa, Aoraki
- Most trigger loan tokens are PN (60%) or SC (28%)
- Shorter n-grams more common (83% = 1 word, 11% = 2 words)
- 21% core, 19% cultural
- 80% listed in dictionary

Background Data & Methods Findings Future Work

17

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?



Background Data & Methods Findings Future Work

18

Wrapping Up

- We propose a new method for studying loanwords at a macro-discourse level, called *collo-textualisation*
 - 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?)

19



Thank you!

Special thanks to:

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



Sally Harper



Katie Levendis

20