1

Background

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

...where we are starting from

2

Māori Loanwords – *intensely studied (1991-2019)*

THE UNIVERSITY OF WAIKATO Te Whare Wānanga o Waikato

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

Source: Macalister, J. (2008) The Māori Presence in the English Lexicon, 1850-2000. *English World-Wide*, 27: 1-24.

Loanword use

speaker/write

Loanword use – both in term

Some loanwor native English

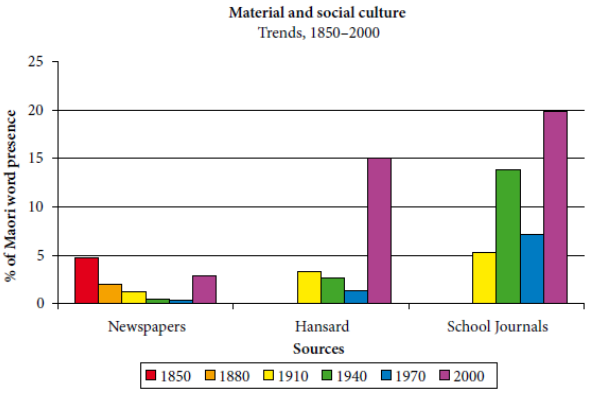


Figure 11. Maori cultural word presence, 1850–2000

3

Māori Loanwords projects

THE UNIVERSITY OF WAIKATO Te Whare Wānanga o Waikato

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)

4

Different data

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

CHEAP TO GET LOTS OF IT — BUT NOISY

FORMAL & INFORMAL
NOT EDITED
CREATIVE
SINGLE-AUTHORED
NORMATIVE & NON-NORM

5

Twitter

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

Twitter
% of internet users who use Twitter

		Use Twitter
All internet users (n=1,802)		**16%**
a	Men (n=846)	17
b	Women (n=956)	15
Race/ethnicity		
a	White, Non-Hispanic (n=1,332)	14
b	Black, Non-Hispanic (n=178)	26[a]
c	Hispanic (n=154)	19
Age		
a	18-29 (n=318)	27[bcd]
b	30-49 (n=532)	16[cd]
c	50-64 (n=551)	10[d]
d	65+ (n=368)	2
Education attainment		
a	Less than high school/high school grad (n=549)	15
b	Some College (n=519)	17
c	College + (n=721)	15

Source: Duggan, M., & Brenner, J. (2013). *The demographics of social media users, 2012* (Vol. 14). Washington, DC: Pew Research Center's Internet & American Life Project.

6

What we did last summer

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

...where we detail corpus building strategy and rationale

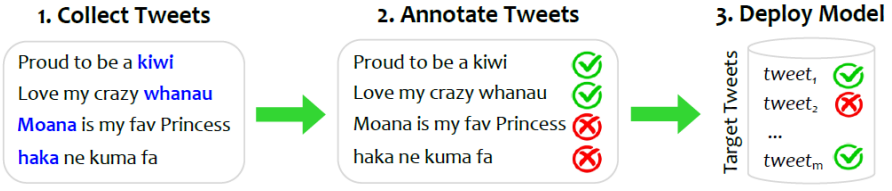
7

Overview

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

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:



8

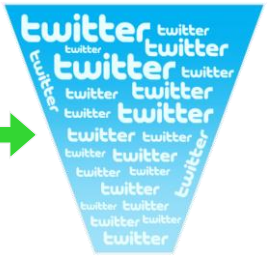
Building the *MLT Corpus* (i)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

116 Query Words

Aotearoa
Aroha
Atua
Awa
⋮
Whero

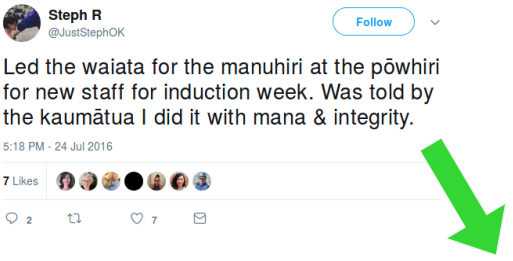
8 Million Tweets



9

Twitter Data

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato



tweet id	username	timestamp (GMT)	query word	text
757369343642480640	JustStephOK	2016-07-25 12:18	***waiata***	Led the ***waiata*** for the manuhiri at the pōwhiri for new staff for induction week. Was told by the kaumātua I did it with mana & integrity.

10

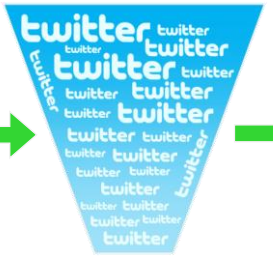
Building the *MLT Corpus* (i)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

116 Query Words

Aotearoa
Aroha
Atua
Awa
⋮
Whero

8 million Tweets



4,600 Labelled Tweets

Proud to be a *kiwi* ✓
Love my crazy *whanau* ✓
Moana is my fav Princess ✗
haka ne kuma fa ✗

11

Common Types of Noise

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

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

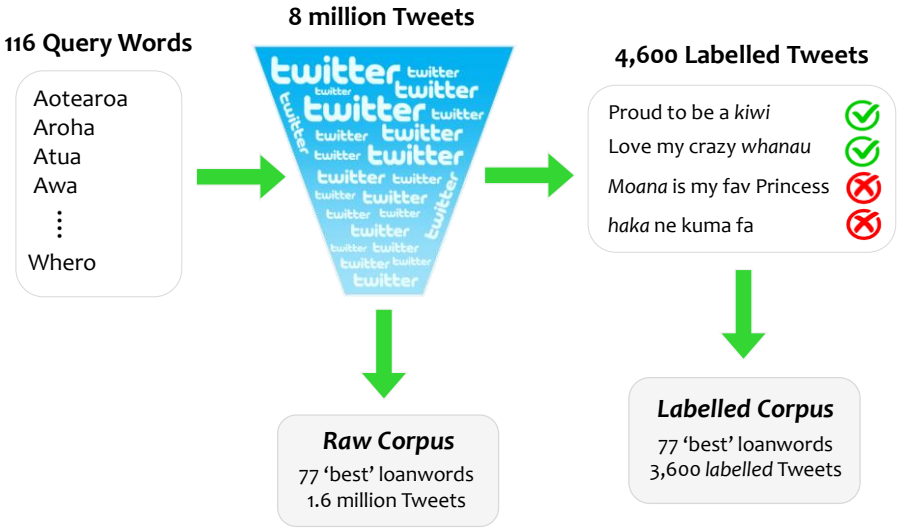
2) Proper Nouns
- Personal/place name used (rather than content word)
- Theoretically count as loanwords, but their use does not constitute a choice
- e.g. *Moana* used to refer to Disney princess/film

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"

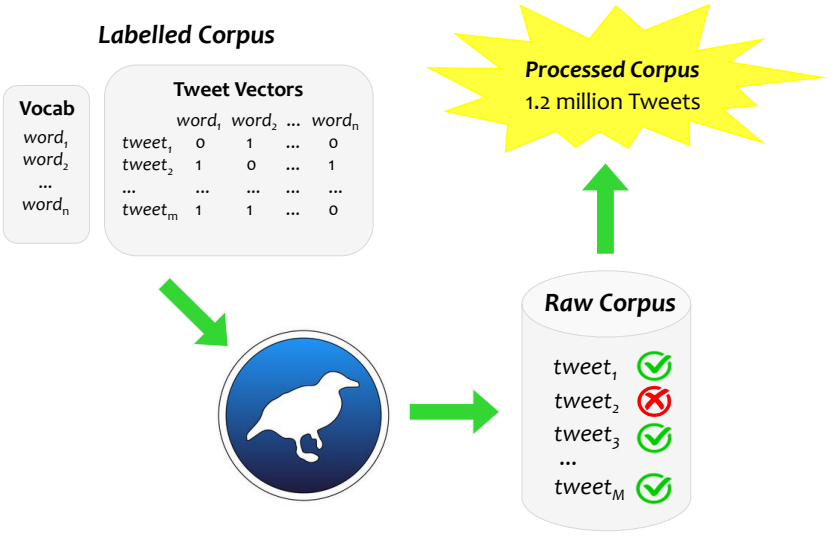
12

Building the *MLT Corpus* (i)



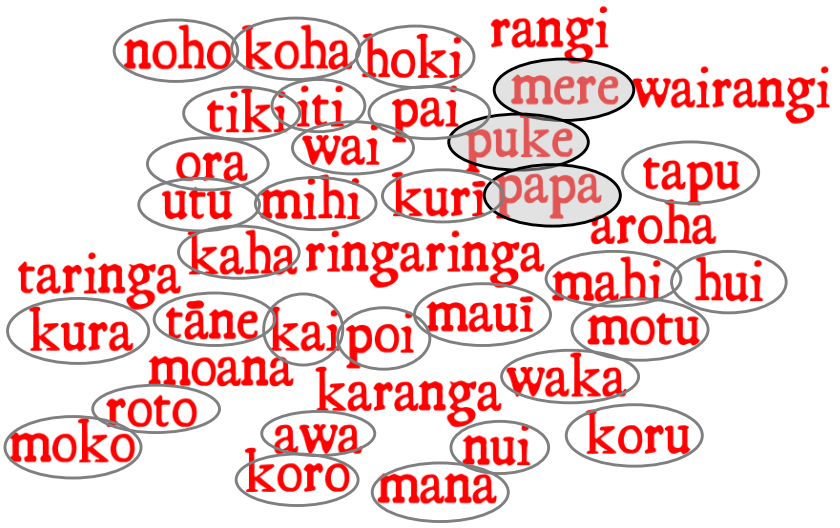
13

Building the *MLT Corpus* (ii)



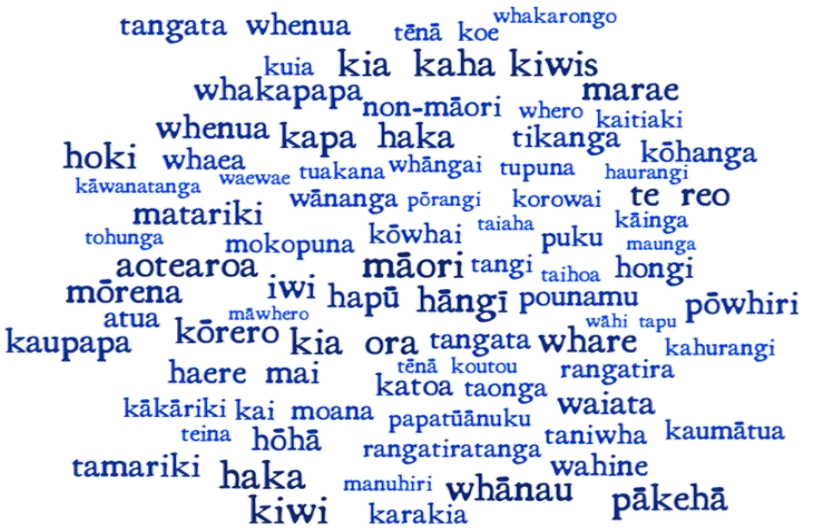
14

Discarded Query Words



15

77 'Best' Query Words



16

Machine Learning Input and Output

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Training Data (Input)

Id	username	timestamp	query word	text	relevance label
7573693436424840640	JustStephOK	2016-07-25 12:18	***waiata***	Led the ***waiata*** for the manuhiri at the pōwhiri for new staff for induction week. Was told by the kaumātua I did it with mana & integrity.	Relevant

Target Data (Output)

id	username	timestamp	query word	text	prob_rel
80958924403756 6460	KUOI_DJ	2016-12-16 15:41	***waiata***	Split Enz—History Never Repeats— ***Waiata***	0.078 (irrelevant)

Classification Results

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

	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

Corpus Statistics

	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

19

What we are doing now...

… where we discuss current analyses

20

Hybrid hashtags

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Jeanette King liked

Leonie Hayden @sharkpatu · 24 Jan 2018
Our review of Paul Moon's book is up! I hope he likes it. #goodtereostories

> **The Spinoff Ātea** @SpinoffAtea
> I pānui a @HemiKelly i te pukapuka hou e whakatutū nei i te puehu nā te tohunga hītōria Pākehā, nā Paul Moon kia kore ai koe e mate ki te pānui. #goodtereostories thespinoff.co.nz/atea/25-01-201…

1 8 38

Wairangi Jones @wairangi58 · 22 Jan 2018
Weell nw Mike,@ #goodtereostories, kura kaupapa, MTV, etc Te Reo is thriving. Guts, persistence, aroha 4 Te Reo means a kai a te kuri like you will neva beat us down!! newstalkzb.co.nz/on-air/mike-ho…

1 5

Show this thread

Te Taura Whiri i te Reo Māori liked

Mighty Ape NZ @MightyApe · 19 Jan 2018
Kia ora koutou, here's some lovely feedback that we have received from a customer about our efforts to incorporate te reo Māori into our communications #LetsShareGoodTeReoStories #GoodTeReoStories

21

Hashtags – what are they?

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

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

(cf. Caleffi 2015)

Chris Messina
(US actor and director)

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

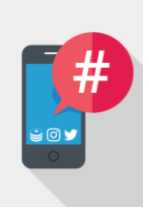
"paralanguage of Twitter" (Maity, Saraf and Mukherjee, 2016:1)

"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)



22

Hashtags – how do they arise?

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

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

(Romero et al 2011)

TOPICS: Association with hashtags with controversial politics and sport makes them more sticky and more pervasive

'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)

Hybrid hashtags (n=38) labelled corpus (45K words)

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

HYBRID
#hakafit
#kiwias
#kaitoputinmyfridge
#maoriwordoftheday
#kaitime
#nethui
#TeReoForLala
#aotearoalove
#māoripotential
#maorizone
#kiwibling
#homeofhaka
#MaoriArt
#MāoriLifeProbs
#transwhanau
#MaoriLanguageWeek /#Maor
#MatarikiTweet
#AotearoaReggaeAllStars
#ChiefsMana
#engagingmāorilearnersconf2(
#maoritv / #MaoriTV
#homewhanau
#WaitangiDay / # waitangiDay
#TearawaStorytellers
#KiwiBuild
#rāmereshorts
#RoosKiwis
#BringitonMana
#keepinItReo
#GoKiwis
#VoteMarama
#FindMeAMaoriBride
#growingupkiwi
#SaveOurKauri
#kiwichicksrock
#learntereo
#LetsShareGoodTeReoStories
#SnortAllThePēpi

25

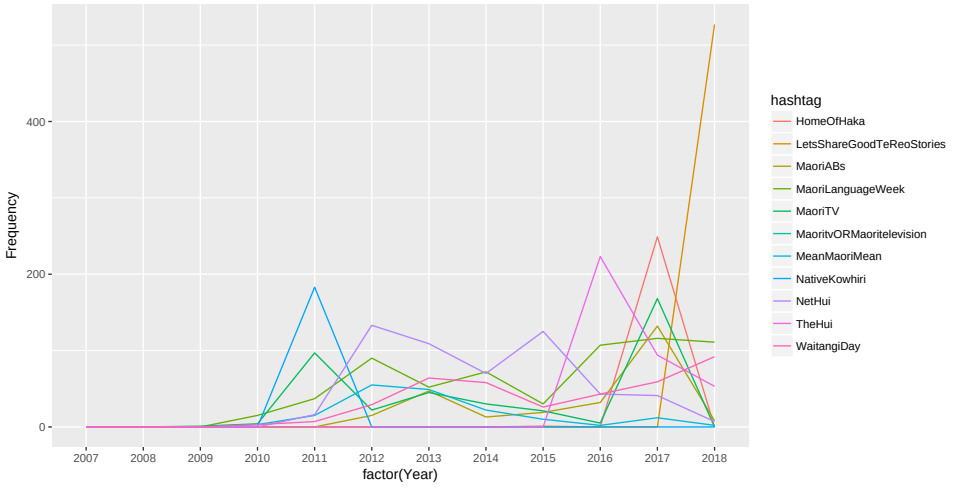
Hybrid hashtags (n=60) processed corpus freq>15

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

hashtag	type	g_hash(
#MaoriLanguageWeek	hybrid	630
#NetHui	hybrid	544
#letssharegoodtereostories	hybrid	527
#MaoritvORMaoritelevision	hybrid	393
#WaitangiDay	hybrid	381
#thehui	hybrid	370
#MaoriABs	hybrid	266
#HomeOfHaka	hybrid	250
#NativeKowhiri	hybrid	183
#MeanMaoriMean	hybrid	170
#Waatea5thEstate	hybrid	164
#ProudKiwi	hybrid	158
#ChiefsMana_CHIEFSMANA_Chie	hybrid	130
#GrowingUpKiwi	hybrid	103
#beingmāoriORbeingmaori	hybrid	81
#kiwipride	hybrid	79
#kiwisongs	hybrid	73
#PakehaParty	hybrid	70
#hakarena	hybrid	69
#FindMeAMaoriBrideORwithmacro	hybrid	67
#youknowyourekiwiwhen	hybrid	66
#treatyofwaitangi	hybrid	65
#rāmereshorts	hybrid	61
#maoriparty	hybrid	57
#ManaParty	hybrid	55
#gokiwisORGoKiwis	hybrid	52
#proudtobekiwiORproudtobeAkiwi	hybrid	37

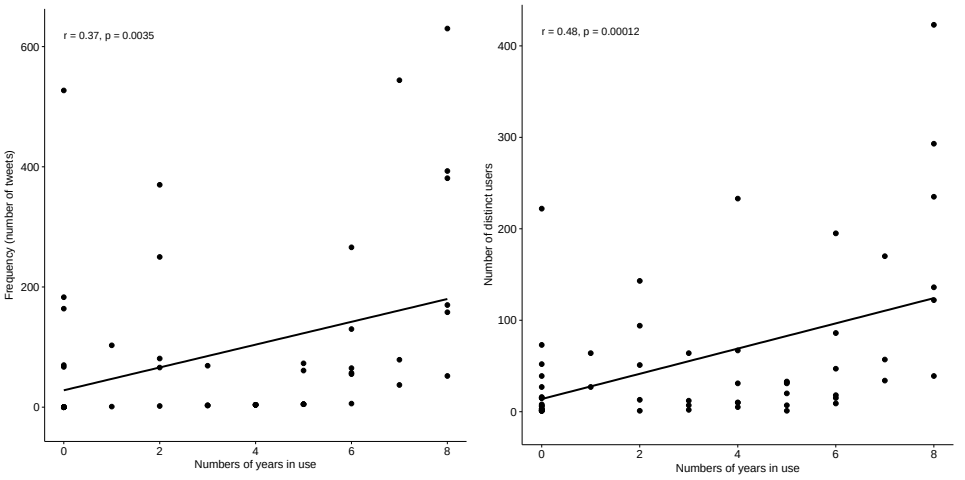
26

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



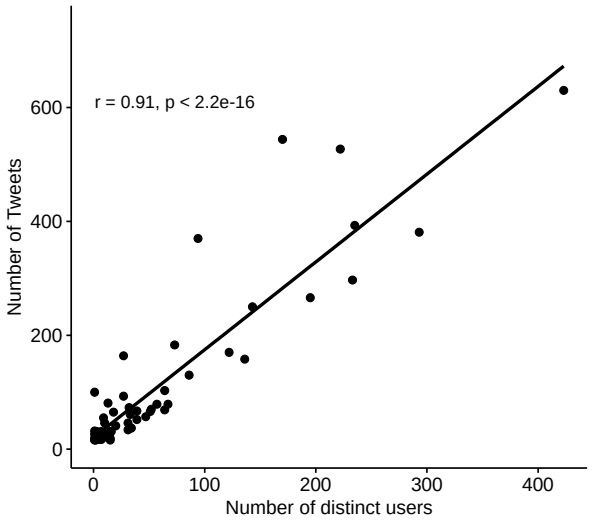
27

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



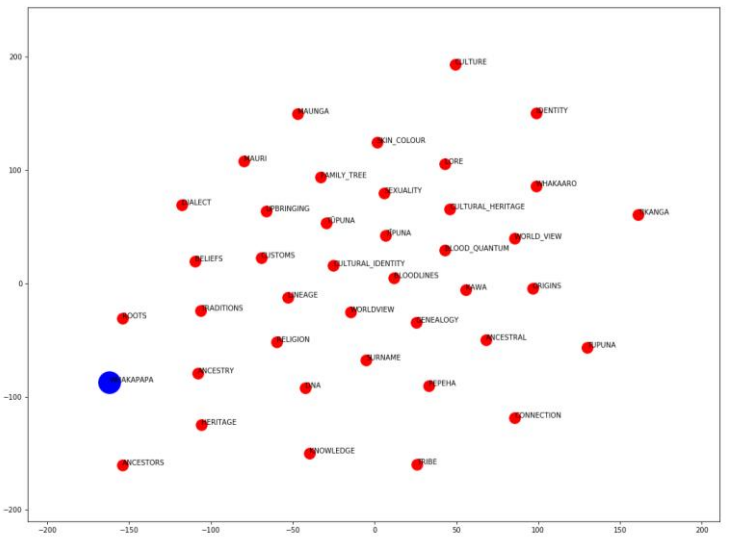
28

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



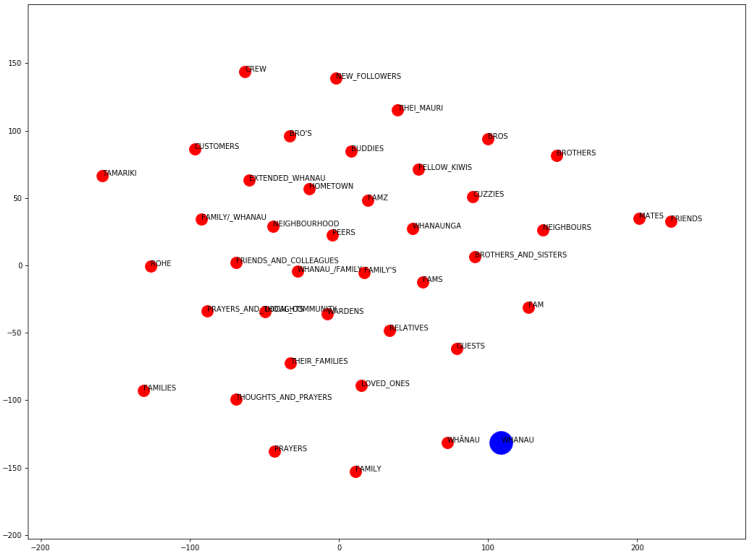
29

Whakapapa



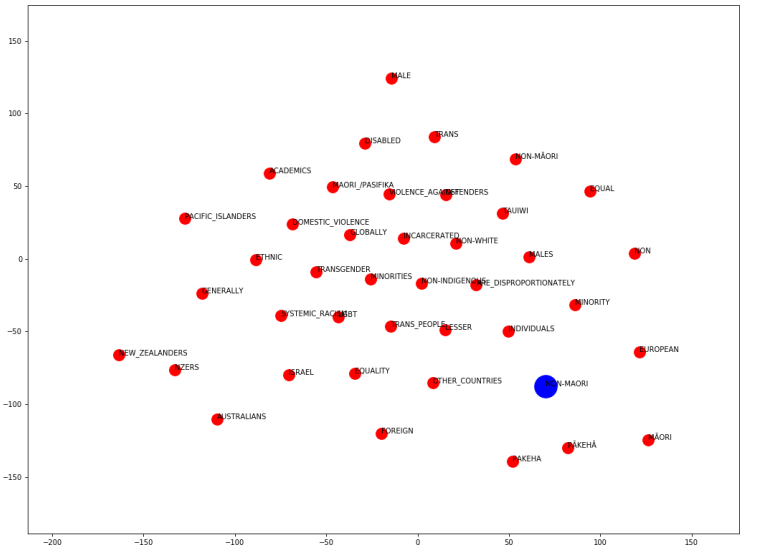
30

Whānau



31

Māori



32

Acknowledgements

Tēnā koutou katoa!

MARSDEN FUND
TE PŪTEA RANGAHAU A MARSDEN
the ROYAL SOCIETY of NEW ZEALAND
TE APĀRANGI
KO TE TANGATA
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato