Māori Loanwords: A Corpus of New Zealand English Tweets

David Trye
Computing and Mathematical Sciences
University of Waikato, New Zealand
dgt12@students.waikato.ac.nz

Andreea S. Calude
School of General and Applied Linguistics
University of Waikato, New Zealand
andreea.calude@waikato.ac.nz

Felipe Bravo-Marquez
Department of Computer Science
University of Chile & IMFD
fbravo@dcc.uchile.cl

Te Taka Keegan
Computing and Mathematical Sciences
University of Waikato, New Zealand
tetaka.keegan@waikato.ac.nz

Abstract

Māori loanwords are widely used in New Zealand English for various social functions by New Zealanders within and outside of the Māori community. Motivated by the lack of linguistic resources for studying how Māori loanwords are used in social media, we present a new corpus of New Zealand English tweets. We collected tweets containing selected Māori words that are likely to be known by New Zealanders who do not speak Māori. Since over 30% of these words turned out to be irrelevant (e.g., *mana* is a popular gaming term, *Moana* is a character from a Disney movie), we manually annotated a sample of our tweets into relevant and irrelevant categories. This data was used to train machine learning models to automatically filter out irrelevant tweets.

1 Introduction

One of the most salient features of New Zealand English (NZE) is the widespread use of Māori words (loanwords), such as *aroha* (love), *kai* (food) and *Aotearoa* (New Zealand). See ex. (1) specifically from Twitter (note the informal, conversational style and the Māori loanwords emphasised in bold).

(1) 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** and integrity.

The use of Māori words has been studied intensively over the past thirty years, offering a comprehensive insight into the evolution of one of the youngest dialects of English – New Zealand English (Calude et al., 2017; Daly, 2007, 2016; Davies and Maclagan, 2006; De Bres, 2006; Degani and Onysko, 2010; Kennedy and Yamazaki, 1999; Macalister, 2009, 2006a; Onysko and Calude, 2013). One aspect which is missing in this body of work is the online discourse presence of the loanwords - almost all studies come from (collaborative) written language (highly edited, revised and scrutinised newspaper language, Davies and Maclagan 2006; Macalister 2009, 2006a,b; Onysko and Calude 2013, and picture-books, Daly 2007, 2016), or from spoken language collected in the late 1990s (Kennedy and Yamazaki, 1999).

In this paper, we build a corpus of New Zealand English tweets containing Māori loanwords. Building such a corpus has its challenges (as discussed in Section 3.1). Before we discuss these, it is important to highlight the uniqueness of the language contact situation between Māori and (NZ) English.

The language contact situation in New Zealand provides a unique case-study for loanwords because of a number of factors. We list three particularly relevant here. First, the direction of lexical transfer is highly unusual, namely, from an endangered indigenous language (Māori) into a dominant lingua franca (English). The large-scale lexical transfer of this type has virtually never been documented elsewhere, to the best of our knowledge (see summary of current language contact situations in Stammers and Deuchar 2012, particularly Table 1, p. 634).

Secondly, because Māori loanwords are "New Zealand's and New Zealand's alone" (Deverson, 1991, p. 18-19), and above speakers' consciousness, their ardent study over the years provides a fruitful comparison of the use of loanwords across genres, contexts and time.

Finally, the aforementioned body of previous research on the topic is rich and detailed, and still rapidly changing, with loanword use being an increasing trend (Macalister, 2006a; Kennedy and Yamazaki, 1999). However, the jury is still out regarding the reasons for the loanword use (some hypotheses have been put forward), and the pat-

terns of use across different genres (it is unclear how language formality influences loanword use).

We find that Twitter data complements the growing body of work on Māori loanwords in NZE, by adding a combination of institutional and individual linguistic exchanges, in a non-editable online platform. Social media language shares properties with both spoken and written language, but is not exactly like either. More specifically, Twitter allows for creative expression and lexical innovation (Grieve et al., 2017).

Our Twitter corpus was created by following three main steps: collecting tweets over a ten-year period using "query words" (Section 3.1), manually labelling thousands of randomly-sampled tweets as "relevant" or "irrelevant" (Section 3.2), and then training a classifier to obtain automatic predictions for the relevance of each tweet and deploying this model on our target tweets, in a bid to filter out all those which are "irrelevant" (Section 3.3). As will be discussed in Section 2, our corpus is not the first of its kind but is the first corpus of New Zealand English tweets and the first collection of online discourse built specifically to analyse the use of Māori loanwords in NZE. Section 4 outlines some preliminary findings from our corpus and Section 5 lays out plans for future work.

2 Related Work

It is uncontroversial that Māori loanwords are both productively used in NZE and increasing in popularity (Macalister, 2006a). The corpora analysed previously indicate that loanword use is highly skewed, with some language users leading the way – specifically Māori women (Calude et al., 2017; Kennedy and Yamazaki, 1999), and with certain topics of discourse drawing significantly higher counts of loanwords than others – specifically those related to Māori people and Māori affairs, *Māoritanga* (Degani, 2010). The type of loanwords being borrowed from Māori is also changing. During the first wave of borrowing, some two-hundred years ago, many flora and fauna words were being borrowed; today, it is social culture terms that are increasingly adopted, e.g., *aroha* (love), *whaea* (woman, teacher), and *tangi* (Māori funeral), see Macalister (2006a). However, the data available for loanword analysis is either outdated (Calude et al., 2017; Kennedy and Yamazaki, 1999), or exclusively formal and highly edited (mainly newspaper language, Macalister 2006a; Davies and Maclagan 2006; Degani 2010), so little is known about Māori loanwords in recent informal NZE interactions – a gap we hope to address here.

With the availability of vast amounts of data, building Twitter corpora has been a fruitful endeavour in various languages, including Turkish (Şimşek and Özdemir, 2012; Çetinoglu, 2016), Greek (Sifianou, 2015), German (Scheffler, 2014; Cieliebak et al., 2017), and (American) English (Huang et al., 2016) (though notably, not New Zealand English, while a modest corpus of te reo Māori tweets does exist, Keegan et al. 2015). Twitter corpora of mixed languages are tougher to collect because it is not straightforward to detect mixed language data automatically. Geolocations can help to some extent, but they have limitations (most users do not use them to begin with). Recent work on Arabic has leveraged the presence of distinct scripts – the Roman and Arabic alphabet – to create a mixed language corpus (Voss et al., 2014), but this option is not available to us. Māori has traditionally been a spoken (only) language, and was first written down in the early 1800s by European missionaries in conjunction with Māori language scholars, using the Roman alphabet (Smyth, 1946). Our task is more similar to studies such as Das and Gambäck (2014) and Çetinoglu (2016), who aim to find a mix of two languages which share the same script (in their case, Hindi and English, and Turkish and German, respectively), but our method for collecting tweets is not user-based; instead we use a set of target query words, as detailed in Section 3.1.

3 The Corpus

In this section, we describe the process of building the Māori Loanword Twitter Corpus (hereafter, the *MLT Corpus*)[1]. This process consists of three main steps, as depicted in Figure 1.

3.1 Step 1: Collecting Tweets

In order to facilitate the collection of relevant data for the *MLT Corpus*, we compiled a list of 116 target loanwords, which we will call "query words".

[1]The corpus is available online at https://kiwiwords.cms.waikato.ac.nz/corpus/. Note that we have only released the tweet IDs, together with a download script, in accordance with Twitter's terms and conditions. We have also released the list of query words used.

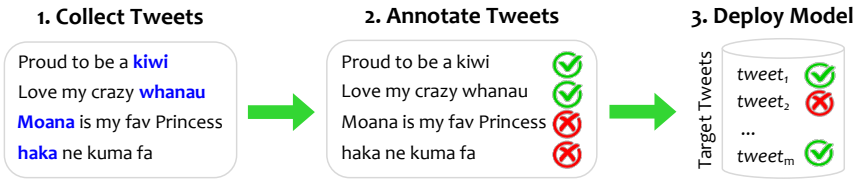

Figure 1: The corpus-building process.

Most of these are individual words but some are short phrasal units (*tangata whenua*, people of the land; *kapa haka*, cultural performance). The list is largely derived from Hay (2018) but was modified to exclude function words (such as numerals) and most proper nouns, except five that have native English counterparts: *Aotearoa* (New Zealand), *Kiwi(s)* (New Zealander(s)), *Māori* (indigenous New Zealander), *Pākehā* (European New Zealander), *non-Māori* (non-indigenous New Zealander). We also added three further loanwords which we deemed useful for increasing our data, namely *haurangi* (drunk), *wairangi* (drugged, confused), and *pōrangi* (crazy).

Using the Twitter Search API, we harvested 8 million tweets containing at least one query word (after converting all characters to lowercase). The tweets were collected diachronically over an eleven year period, between 2007-2018. We ensured that tweets were (mostly) written in English by using the *lang:en* parameter.

A number of exclusions and further adjustments were made. With the aim of avoiding redundancy and uninformative data, retweets and tweets with URLs were discarded. Tweets in which the query word was used as part of a username or mention (e.g., @*happy_kiwi*) were also discarded. For those query words which contained macrons, we found that users were inconsistent in their macron use. Consequently, we consolidated the data by adjusting our search to include both the macron and the non-macron version (e.g., both *Māori* and *Maori*). We also removed all tweets containing fewer than five tokens (words), due to insufficient context of analysis.

Owing to relaxed spelling conventions on Twitter (and also the use of hashtags), certain query words comprising multiple lexical items were stripped of spaces in order to harvest all variants of the phrasal units (e.g., *kai moana* and *kaimoana*). As *kai* was itself a query word (in its own right), we excluded tweets containing *kai moana* when searching for tweets containing *kai* (and repeated this process with similar items).

After inspecting these tweets, it was clear that a large number of our query words were polysemous (or otherwise unrelated to NZE), and had introduced a significant amount of noise into the data. The four main challenges we encountered are described below.

First, because Twitter contains many different varieties of English, NZE being just one of these, it is not always straightforward to disentangle the dialect of English spoken in New Zealand from other dialects of English. This could be a problem when, for instance, a Māori word like *Moana* (sea) is used in American English tweets to denote the Disney movie (or its main character).

Secondly, Māori words have cognate forms with other Austronesian languages, such as Hawaiian, Samoan and Tongan, and many speakers of these languages live and work (and tweet) in New Zealand. For instance, the word *wahine* (woman) has the same written form in Māori and in Hawaiian. But cognates are not the only problematic words. Homographs with other, genealogically-unrelated languages can also pose problems. For instance, the Māori word *hui* (meeting) is sometimes used as a proper noun in Chinese, as can be seen in the following tweet: "Yo is Tay Peng Hui okay with the tip of his finger?".

Proper nouns constitute a third problematic aspect in our data. As is typical for many language contact situations where an indigenous language shares the same geographical space as an incoming language, Māori has contributed many place names and personal names to NZE, such as *Timaru*, *Aoraki*, *Titirangi*, *Hēmi*, *Mere* and so on. While these proper nouns theoretically count as loanwords, we are less interested in them than in content words because the use of the former does not constitute a choice, whereas the use of the latter does (in many cases). The "choice" of whether to use a loanword or whether to use a native English word (or sometimes a native English phrase) is interesting to study because it provides insights

into idiolectal lexical preferences (which words different speakers or writers prefer in given contexts) and relative borrowing success rates (Calude et al., 2017; Zenner et al., 2012).

Finally, given the impromptu and spontaneous nature of Twitter in general, we found that certain Māori words coincided with misspelled versions of intended native English words, e.g., *whare* (house) instead of *where*.

The resulting collection of tweets, termed the *Original Dataset*, was used to create the *Raw Corpus*, as explained below.

3.2 Step 2: Manually Annotating Tweets

We decided to address the "noisy" tweets in our data using supervised machine learning. Two coders manually inspected a random sample of 30 tweets for each query word, by checking the word's context of use, and labelled each tweet as "relevant" or "irrelevant". For example, a tweet like that in example (1) would be coded as relevant and one like " awesome!! Congrats to Tangi :)", would be coded as irrelevant (because the query word *tangi* is used as a proper noun). Since 39 of the query words consistently yielded irrelevant tweets (at least 90% of the time), these (and the tweets they occurred in) were removed altogether from the data. Our annotators produced a total of 3,685 labelled tweets for the remaining 77 query words, which comprise the *Labelled Corpus* (see Tables 1 and 4; note that irrelevant tweets have been removed from the latter for linguistic analysis).

Assuming our coded samples are representative of the real distribution of relevant/irrelevant tweets that occur with each query word, it makes sense to also discard the 39 "noisy" query words from our *Original Dataset*. In this way, we created the (unlabelled) *Raw Corpus*, which is a fifth of the size (see Table 4).

We computed an inter-rater reliability score for our two coders, based on a random sample of 200 tweets. Using Cohen's Kappa, we calculated this value to be 0.87 ("strong"). In light of the strong agreement between the initial coders, no further coders were enlisted for the task.

3.3 Step 3: Automatically Extracting Relevant Tweets

The next step was to train a classifier using the *Labelled Corpus* as training data, so that the resulting model could be deployed on the *Raw Corpus*. Our goal is to obtain automatic predictions for the relevance of each tweet in this corpus, according to probabilities given by our model.

We created (stratified) test and training sets that maintain the same proportion of relevant and irrelevant tweets associated with each query word in the *Labelled Corpus*. We chose to include 80% of these tweets in the training set and 20% in the test set (see Table 1 for a break-down of relevant and irrelevant instances).

	Train	Test	Total
Relevant	1,995	500	2,495
Irrelevant	954	236	1,190
Total	2,949	736	3,685

Table 1: Dataset statistics for our labelled tweets. This Table shows the relevant, irrelevant and total number of instances (i.e., tweets) in the independent training and test sets.

Using the *AffectiveTweets* package (Bravo-Marquez et al., 2019), our labelled tweets were transformed into feature vectors based on the word n-grams they contain. We then trained various classification models on this transformed data in Weka (Hall et al., 2009). The models we tested were 1) Multinomial Naive Bayes (McCallum et al., 1998) with unigram attributes and 2) L2-regularised logistic regression models with different word n-gram features, as implemented in LIBLINEAR[2]. We selected Multinomial Naive Bayes as the best model because it produced the highest AUC, Kappa and weighted average F-Score (see Table 2 for a summary of results). Overall, logistic regression with unigrams performed the worst, yielding (slightly) lower values for all three measures.

After deploying the Multinomial Naive Bayes model on the *Raw Corpus*, we found that 1,179,390 tweets were classified as relevant and 448,652 as irrelevant (with probability threshold = 0.5).

Table 3 shows examples from our corpus of each type of classification. Some tweets were falsely classified as "irrelevant" and some were falsely classified as "relevant". A short explanation why the irrelevant tweets were coded as such is given in brackets at the end of each tweet.

We removed all tweets classified as irrelevant,

[2] https://www.csie.ntu.edu.tw/~cjlin/liblinear/

	AUC	Kappa	F-Score
Multinomial Naive Bayes			
$n = 1$	**0.872**	**0.570**	**0.817**
Logistic Regression			
$n = 1$	0.863	0.534	0.801
$n = 1, 2$	0.868	**0.570**	0.816
$n = 1, 2, 3$	0.869	0.560	0.811
$n = 1, 2, 3, 4$	0.869	0.563	0.813
$n = 1, 2, 3, 4, 5$	0.869	0.556	0.810

Table 2: Classification results on the test set. The best results for each column are shown in **bold**. The value of n corresponds to the type of word n-grams included in the feature space.

thereby producing the *Processed Corpus*. A summary of all three corpora is given in Table 4.

4 Preliminary Findings

As we are only just beginning to sift through the *MLT Corpus*, we note two particular sets of preliminary findings.

First, even though our corpus was primarily geared up to investigate loanword use, we are finding that, unlike other NZE genres analysed, the Twitter data exhibits use of Māori which is more in line with code-switching than with loanword use, see ex. (2-3). This is particularly interesting in light of the reported increase in te reo Māori language tweets (Keegan et al., 2015).

(2) **Mōrena e hoa!** We must really meet IRL when I get back to Tāmaki Makaurau! You have a fab day too!

(3) Heh! **He porangi toku ngeru** - especially at 5 in the morning!! **Ata marie e hoa ma**. I am well thank you.

Secondly, we also report the use of hybrid hashtags, that is, hashtags which contain a Māori part and an English part, for example *#mycrazywhanau, #reostories, #Matarikistar, #bringitonmana, #growingupkiwi, #kaitoputinmyfridge*. To our knowledge, these hybrid hashtags have never been analysed in the current literature. Hybrid hashtags parallel the phenomenon of hybrid compounds discussed by Degani and Onysko (2010). Degani and Onysko report that hybrid compounds are both productive and semantically novel, showing that the borrowed words take on reconceptualised meanings in their adoptive language (2010, p.231).

	Irrelevant tweets	Relevant tweets
f(x)<0.5 Classified irrelevant	**Haka ne!** And i know even the good guys get blood for body (0.282, foreign language)	son didnt get my chop ciggies 2day so stopped talking 2 him. he just walked past and gave me the **maori** eyebrow lift and a smile. were friends (0.337)
	Whare has the year gone (0.36, misspelling)	Shorts and bare feet in this **whare** (0.41)
	chegar na **morena** e falar can i be your girlfriend can i (0.384, foreign language)	**Tangata whenua** charged for killing 6 **#kererū** for **Kai** meanwhile forestry corps kill off widespread habitat for millions #efficiency #doc (0.306)
f(x)≥0.5 Classified relevant	**Te Wanganga o Aotearoa's** moving to a new campus in Palmy, but their media person has refused to talk to us about it. #whatajoke (0.998, proper noun)	Our whole worldview as Maori is **whanau** based. **Pakeha** call it nepotism, tribalism, gangsterism, LinkedInism blah de blah. It's our way of doing stuff and it's not going to change to suit another point of view. (0.995)
	I cant commit to anything but if I were to commit to one song, it would be **kiwi** - harry styles (0.791, proper noun)	**Kia ora koutou** - does anyone know the **te reo** word for Cornwall? (1.0)
	Why am I getting headaches out of no **whero** never get them :(I guess its all the stress (0.542, spelling mistake)	The New Zealand team do another energetic **haka** though (0.956)

Table 3: A selection of tweets and their classification types. The first three irrelevant tweets were classified correctly (i.e. true negatives), as were the last three relevant tweets (i.e. true positives). Function $f(x)$ corresponds to the posterior probability of the "relevant" class. The entries in brackets for the irrelevant examples correspond to the values of $f(x)$ and the reason why the target word was coded as irrelevant.

	Raw	Labelled	Processed
Tokens (words)	28,804,640	49,477	21,810,637
Tweets	1,628,042	2,495	1,179,390
Tweeters (authors)	604,006	1,866	426,280

Table 4: A description of the *MLT Corpus*' three components (namely, the *Raw Corpus*, *Labelled Corpus* and *Processed Corpus*), which were harvested using the same 77 query words.

5 Conclusions and Future Work

This paper introduced the first purpose-built corpus of Māori loanwords on Twitter, as well as a methodology for automatically filtering out irrelevant data via machine learning. The *MLT Corpus* opens up a myriad of opportunities for future work.

Since our corpus is a diachronic one (i.e., all tweets are time-stamped), we are planning to use it for testing hypotheses about language change. This is especially desirable in the context of New Zealand English, which has recently undergone considerable change as it comes into the final stage of dialect formation (Schneider, 2003).

Another avenue of future research is to automatically identify other Māori loanwords that are not part of our initial list of query words. This could

be achieved by deploying a language detector tool on every unique word in the corpus (Martins and Silva, 2005). The "discovered" words could be used as new query words to further expand our corpus.

In addition, we intend to explore the meaning of our Māori loanwords using distributional semantic models. We will train popular word embeddings algorithms on the *MLT Corpus*, such as *Word2Vec* (Mikolov et al., 2013) and *FastText* (Bojanowski et al., 2017), and identify words that are close to our loanwords in the semantic space. We predict that these neighbouring words will enable us to understand the semantic make-up of our loanwords according to their usage.

Finally, we hope to extrapolate these findings by deploying our trained classifier on other online discourse sources, such as *Reddit* posts. This has great potential for enriching our understanding of how Māori loanwords are used in social media.

6 Acknowledgements

The authors would like to thank former Honours student Nicole Chan for a preliminary study on Māori Loanwords in Twitter. Felipe Bravo-Marquez was funded by Millennium Institute for Foundational Research on Data. Andreea S. Calude acknowledges the support of the NZ Royal Society Marsden Grant. David Trye acknowledges the generous support of the Computing and Mathematical Sciences group at the University of Waikato.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Felipe Bravo-Marquez, Eibe Frank, Bernhard Pfahringer, and Saif M. Mohammad. 2019. AffectiveTweets: a Weka package for analyzing affect in tweets. *Journal of Machine Learning Research*, 20:1–6.

Andreea Simona Calude, Steven Miller, and Mark Pagel. 2017. Modelling loanword success–a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*.

Özlem Çetinoglu. 2016. A Turkish-German code-switching corpus. In *International Conference on Language Resources and Evaluation*.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A twitter corpus and benchmark resources for German sentiment analysis. In *5th International Workshop on Natural Language Processing for Social Media, Boston, MA, USA, December 11, 2017*, pages 45–51. Association for Computational Linguistics.

Nicola Daly. 2007. Kūkupa, koro, and kai: The use of Māori vocabulary items in New Zealand English children's picture books.

Nicola Daly. 2016. Dual language picturebooks in English and Māori. *Bookbird: A Journal of International Children's Literature*, 54(3):10–17.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text.

Carolyn Davies and Margaret Maclagan. 2006. Māori words–read all about it: Testing the presence of 13 māori words in four New Zealand newspapers from 1997 to 2004. *Te Reo*, 49.

Julia De Bres. 2006. Maori lexical items in the mainstream television news in New Zealand. *New Zealand English Journal*, 20:17.

Marta Degani. 2010. The Pakeha myth of one New Zealand/Aotearoa: An exploration in the use of Maori loanwords in New Zealand English. *From international to local English–and back again*, pages 165–196.

Marta Degani and Alexander Onysko. 2010. Hybrid compounding in New Zealand English. *World Englishes*, 29(2):209–233.

Tony Deverson. 1991. New Zealand English lexis: the Maori dimension. *English Today*, 7(2):18–25.

Jack Grieve, Andrea Nini, and Diansheng Guo. 2017. Analyzing lexical emergence in Modern American English online. *English Language & Linguistics*, 21(1):99–127.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18.

Jennifer Hay. 2018. What does it mean to "know a word?". In *Language and Society Conference of NZ in November 2018 in Wellington, NZ*, Wellington, NZ.

Yuan Huang, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding US regional linguistic variation with twitter data analysis. *Computers, Environment and Urban Systems*, 59:244–255.

Te Taka Keegan, Paora Mato, and Stacey Ruru. 2015. Using Twitter in an indigenous language: An analysis of Te Reo Māori tweets. *AlterNative: An International Journal of Indigenous Peoples*, 11(1):59–75.

Graeme Kennedy and Shunji Yamazaki. 1999. The influence of Maori on the Nw Zealand English lexicon. *LANGUAGE AND COMPUTERS*, 30:33–44.

John Macalister. 2006a. The Maori lexical presence in New Zealand English: Constructing a corpus for diachronic change. *Corpora*, 1(1):85–98.

John Macalister. 2006b. the Maori presence in the New Zealand English lexicon, 1850–2000: Evidence from a corpus-based study. *English World-Wide*, 27(1):1–24.

John Macalister. 2009. Investigating the changing use of Te Reo. *NZ Words*, 13:3–4.

Bruno Martins and Mário J Silva. 2005. Language identification in web pages. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 764–768. ACM.

Andrew McCallum, Kamal Nigam, et al. 1998. A comparison of event models for naive Bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 752, pages 41–48. Citeseer.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Alexander Onysko and Andreea Calude. 2013. Comparing the usage of Māori loans in spoken and written New Zealand English: A case study of Māori, Pākehā, and Kiwi. *New perspectives on lexical borrowing: Onomasiological, methodological, and phraseological innovations*, pages 143–170.

Tatjana Scheffler. 2014. A German twitter snapshot. In *LREC*, pages 2284–2289. Citeseer.

Edgar W Schneider. 2003. The dynamics of New Englishes: From identity construction to dialect birth. *Language*, 79(2):233–281.

Maria Sifianou. 2015. Conceptualizing politeness in Greek: Evidence from twitter corpora. *Journal of Pragmatics*, 86:25–30.

Mehmet Ulvi Şimşek and Suat Özdemir. 2012. Analysis of the relation between Turkish twitter messages and stock market index. In *2012 6th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4. IEEE.

Patrick Smyth. 1946. *Maori Pronunciation and the Evolution of Written Maori*. Whitcombe & Tombs Limited.

Jonathan R Stammers and Margaret Deuchar. 2012. Testing the nonce borrowing hypothesis: Counterevidence from English-origin verbs in Welsh. *Bilingualism: Language and Cognition*, 15(3):630–643.

Clare R Voss, Stephen Tratz, Jamal Laoudi, and Douglas M Briesch. 2014. Finding Romanized Arabic dialect in code-mixed tweets. In *LREC*, pages 2249–2253.

Eline Zenner, Dirk Speelman, and Dirk Geeraerts. 2012. Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch. *Cognitive Linguistics*, 23(4):749–792.