When loanwords are not lone words: Using networks and hypergraphs to explore Māori loanwords in New Zealand English

David Trye, Andreea S. Calude, Te Taka Keegan, Julia Falconer

DOI: https://doi.org/10.1075/ijcl.21124.try
To appear in: International Journal of Corpus Linguistics

Received date: 26 August 2021
Accepted date: 15 July 2022

Please cite this article as: Trye, D., Calude, A. S., Keegan, T. T., & Falconer, J. (2023). When loanwords are not lone words: Using networks and hypergraphs to explore Māori loanwords in New Zealand English. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.21124.try.

When loanwords are not lone words

Using networks and hypergraphs to explore Māori loanwords in New Zealand English

Networks are being used to model an increasingly diverse range of real-world phenomena. This paper introduces an exploratory approach to studying loanwords in relation to one another, using networks of co-occurrence. While traditional studies treat individual loanwords as discrete items, we show that insights can be gained by focusing on the various loanwords that co-occur within each text in a corpus, especially when leveraging the notion of a hypergraph. Our research involves a case-study of New Zealand English (NZE), which borrows Indigenous Māori words on a large scale. We use a topic-constrained corpus to show that: (i) Māori loanword types tend not to occur by themselves in a text; (ii) infrequent loanwords are nearly always accompanied by frequent loanwords; and (iii) it is not uncommon for texts to contain a mixture of listed and unlisted loanwords, suggesting that NZE is still riding a wave of borrowing importation from Māori.

Keywords: loanwords, networks, hypergraphs, New Zealand English, Māori

1. Introduction

As we live in an evermore-connected world, the effects of bilingualism and multilingualism are becoming increasingly salient, even as far as monolingual speakers are concerned. The most obvious and pervasive effect observed in situations of language contact is the borrowing of words from one language into another. Loanwords, also called borrowings (we use these terms interchangeably), have preoccupied linguists for nearly a century (Haugen, 1950; Weinreich, 1953).

Unsurprisingly, most loanword studies document situations of language contact in which English words are adopted into other languages (e.g. see Görlach, 2002 for a discussion of Anglicisms in European language). The current study focuses on a different direction of borrowing, namely on words borrowed *into* English from Māori, the Indigenous language of Aotearoa[1]. The language contact situation in New Zealand is particularly unusual. This is because words from a non-dominant language undergoing revitalisation (Māori) are being adopted on a large scale by a world-dominating lingua franca: namely, a variety of English called New Zealand English (hereafter, NZE; see Section 2.2).

Given the large body of work focusing on loanwords, a number of different avenues exist for studying their use. Here, we propose a new method for investigating the use of loanwords in a corpus by adopting a macro-discourse framework that considers the co-

occurrence of terms within the same texts, facilitated by network analysis tools. While we probe data from a case-study of NZE, our aim is to present novel quantitative ways of studying loanwords that have wider methodological implications beyond NZE.

Our discourse-oriented approach involves examining loanwords by considering each corpus text as a whole (see Section 3), and extracting loanwords that co-occur in the text as a 'set', rather than listing them as discrete elements. The rationale for this method comes from the observation that NZE texts (in the loose sense of a conversation, newspaper article, etc.) tend to either exhibit several Māori loanwords or none at all. This has been anecdotally noted in children's picture books (Macdonald & Daly, 2013: 48). Moreover, in loanword-rich texts, borrowed items may be dispersed throughout the text, rather than appearing within a small window of one another. In this sense, they do not behave like collocates (Firth, 1957; see Kurtböke & Potter, 2000 for an analysis of loanword collocates) because there is no optimal (fixed) window-size for capturing their co-occurrence. This serves as motivation for changing the window-size based on the position of keywords and the total number of words in the corresponding text.

Our aim is to explore loanwords from this fresh perspective by answering the following questions:

i. How might loanword networks and hypergraphs (Section 4.3) be operationalised using a discourse-oriented approach?
ii. What can studying loanwords by means of networks and hypergraphs tell us about the borrowing process in general?

2. Background

This section begins with a brief overview of the field of loanword research, paying special attention to widely used measures of entrenchment, such as frequency and dispersion (Section 2.1). Additional background information about the language contact situation in New Zealand is then given, together with a summary of related work (Section 2.2), as this provides necessary context for understanding our case-study.

2.1 Entrenchment: What to count, how to count it and what it can tell us

Loanword research has a long and rich history, with scholars studying the transfer of words from one (donor) language into another (receiver) language. This body of work aims to answer

a wide range of questions, from identifying which words a specific language might borrow, to why speakers borrow them in the first place, to how we distinguish between borrowing and related phenomena, such as code-switching. Due to the breadth of the field of loanword research, we limit this section to a summary of key findings and ideas that are especially relevant to this work.

The most common measure employed to capture loanword patterns is frequency of use. More recently, studies have modelled relative loanword success, detailing the lexical competition arising when an incoming loanword encroaches on the semantic space of an existing word in the receiver language (Zenner et al., 2012; Author, 2017). In such studies, the loanwords of interest are typically examined independently, without taking into account any other loanwords that may be present in the same text.

Frequency of use can be helpful for ranking loanwords according to their overall salience and stability. However, difficulties emerge when distinguishing loanwords from code-switches, and it remains unclear whether such a distinction is theoretically warranted (see Poplack, 2018 for a position that favours a strong dichotomy between the two and Stammers & Deuchar, 2012 for a position against it). It should also be noted that, with few exceptions (e.g. Zenner et al., 2013, 2015), loanwords are generally considered to be single lexical items, while 'multiword stretches' are code-switches (Poplack, 2018: 7). As will be discussed in Section 3.1, this is problematic for studying loanwords in NZE, so we consider loan-phrases alongside individual loanwords.

Loanwords can be classified in various ways, depending on their meaning and type, and their use can be tracked diachronically. Such classifications are useful for determining general trends that operate in the receiver language. But how can we be sure that a loanword has successfully and decidedly entered the lexical inventory of a receiver language? Frequency is also enlisted here as an indicator of entrenchment. Most loanwords constitute single-use borrowings (nonce loanwords); in other words, they are only fleeting encounters, further complicating the boundary between loanword use and code-switching. In contrast, recurrent loanwords are highly likely to become integrated in the lexicon of a receiver language. For example, the French word *café* takes the English plural morpheme *-s*: *cafés* (often written without the accent over the *e*). Bilinguals may be the source of loanwords but, ultimately, their success depends on monolinguals adopting them.

Apart from frequency, another indicator of entrenchment is dispersion, also termed 'diffusion' or 'burstiness' (see Chesley & Baayen, 2014; Poplack, 2018, Chapter 4). Like other parts of the lexicon (Zipf, 1935), loanword frequency varies across lexical items, with

some loanwords being generally more frequent than others. While we know that dispersion applies to various parts of the lexicon, it is not straightforward to operationalise (Gries, 2013, 2021). In regard to Māori borrowings in NZE, it has been shown that the topic of discourse influences the use of loanwords (Degani, 2010; Author, 2019), such that Māori-related topics elicit higher counts.

Another factor relevant to entrenchment is acceptance or listedness (see Section 3.4). Discussing code-mixing, Muskyen (2000: 71) distinguishes between what he terms 'creative' use of lexical forms versus 'reproductive' use, which vary in regard to "the degree to which a particular element or structure is part of a memorised list which has gained acceptance within a particular speech community". This distinction is operationalised by Stammers & Deuchar (2012: 631) for English verbs borrowed into Welsh, by verifying the listedness of these verbs in Welsh dictionaries. According to Stammers & Deuchar (ibid), such entries show loanwords that are 'established borrowings'.

2.2 Māori Loanwords in New Zealand English

The data we present here comes from a case-study of Māori loanwords adopted into NZE. Hence, before describing our data and methods, some context about the language contact situation in New Zealand is in order. The language of the Māori people was spoken on the shores of Aotearoa when colonial English settlers first arrived. However, the language these settlers brought with them would become a world lingua franca, and eventually take over as a dominant language in Aotearoa, threatening the vitality of the local Indigenous language.

As a settler colonial variety (Denis & D'Arcy, 2018), NZE has undergone two major 'waves' of lexical borrowing from Māori. The first wave (also called the 'colonisation phase') took place during the initial contact period between Māori and English, upon the arrival of Captain Cook in the late 18th century (Macalister, 2006: 18ff.). This first wave was characterised by borrowings related to the local environment, including words for local flora and fauna (e.g. *kumara* "sweet potato", *manuka* "tea-tree") and various proper nouns (e.g. *Hēmi* "James", *Aotearoa* "New Zealand"). According to Macalister (ibid), the first wave lasted until around 1880, and was followed by a period of resistance to borrowing from 1880-1970. The second wave, the so-called 'decolonisation phase', began shortly thereafter, with a shift towards the borrowing of social and material loanwords (e.g. *kaitiakitanga* "guardianship", *rohe* "tribal boundary").

The use of Māori loanwords in NZE has been studied extensively in various genres, including newspaper articles, spontaneous conversation, online discourse, Twitter data and children's picture books (see Author, 2019 for a comprehensive summary). These studies show widespread, productive and ongoing use of words of Māori origin in NZE, at a normalised rate of six or seven per thousand words (Kennedy, 2001; Macalister, 2006). The majority of studies compute frequency counts for individual loanwords (e.g. Macalister, 2000, 2006, 2009; Davies & Maclagan, 2006; de Bres, 2006; Author, 2019; Author, 2020) or their relative success (Author, 2017). To our knowledge, this is the first attempt to operationalise a method for analysing a large loanword dataset by focusing on the presence of other loanwords in the same text, not just in the NZE context, but in any language contact situation.

3. Methodology

In this section, we describe the data and methods that were used to analyse loanword co-occurrence in NZE. Code for extracting and processing the data can be found at https://github.com/Waikato/kiwiwords/tree/master/loanword_networks. Section 3.1 details the corpus used, and Section 3.2 explains our criteria for selecting loanwords. We then explain how we computed loanword co-occurrence (Section 3.3), outline three linguistic properties of interest (Section 3.4), and provide an overview of the loanwords' overall frequency in the corpus (Section 3.5).

3.1 Overview of the Matariki Corpus

This study investigates loanword co-occurrence within an existing corpus of NZE newspaper articles, called the *Matariki Corpus* (Author, 2019). The corpus was designed to study Māori loanword use by capturing texts that explicitly mention 'Matariki', the Māori New Year, which celebrates the rising of the Pleiades star cluster in late June or early July of each year. As the data consists of newspaper articles, the language used in the Matariki Corpus is planned and edited. Summary statistics for the Matariki Corpus, including its diachronic dimension, are given in Table 1. The corpus has a high loanword rate, likely because the topic of discourse is directly relevant to Māori.

Table 1. Basic summary statistics for the Matariki Corpus

Timeframe	2007-2016
Tokens	91,958
Texts	194
Average Tokens per Text	474
Loanwords per 1,000 words	29

3.2 Loanword Selection Process

The method used for identifying loanwords in the Matariki Corpus involved a combination of computational and manual techniques, summarised in Figure 1.

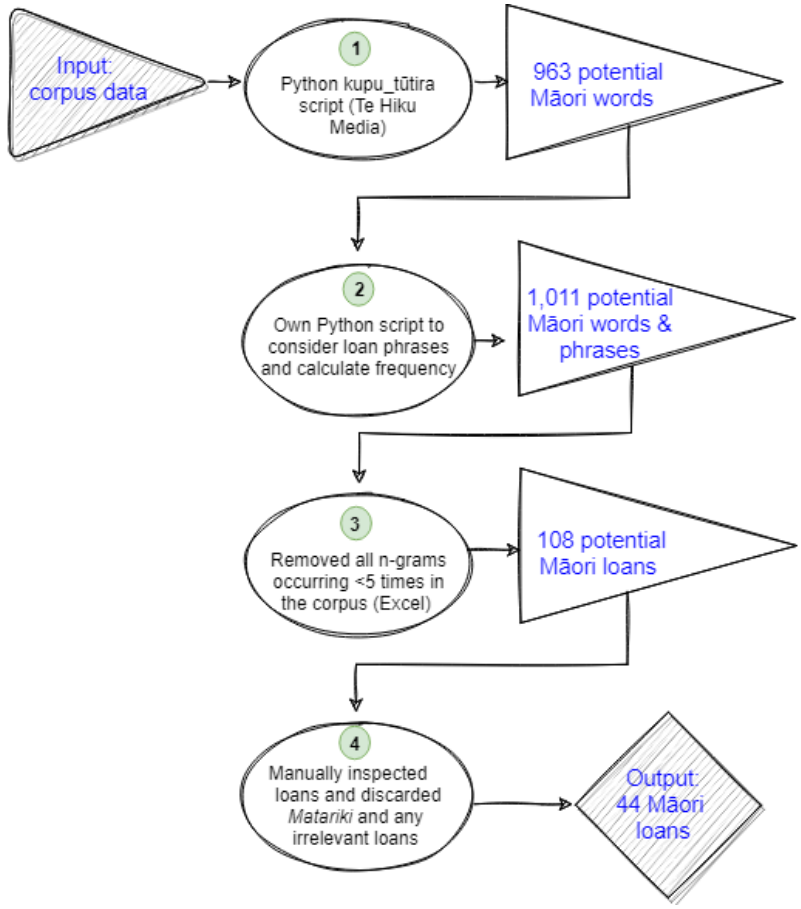

Figure 1. The loanword selection process

The **first step** was to identify potential Māori loanwords in the corpus, by leveraging code developed by Te Hiku Media[2]. The *kupu_tūtira* script was used to obtain a list of words that are consistent with Māori orthography. Such items: (i) consist solely of characters in the Māori alphabet[3]; (ii) follow consonant/vowel alternation; (iii) do not contain double consonants; and (iv) end in a vowel. The resulting wordlist contained 963 items but suffered from two main issues: (i) it did not contain *only* Māori loanwords, because some irrelevant/non-Māori words happened to meet the above criteria (e.g. *make*), and (ii) some of the loanwords occurred in collocate loan-phrases rather than single loan words (e.g. *tangata whenua* "people of the land"). These problems were addressed in steps two-four.

A script was then developed to identify and count collocate Māori words and loan-phrases (**step two**). This process increased our list from 963 items to 1,011 items, because some terms remained free-standing loanwords (e.g. *whenua* "land") but also occurred in loan-phrases (e.g. *tangata whenua*). Although most studies assume loanwords to be single words (see Section 2.1), in our data, Māori loans sometimes occur as multi-word phrases, whose component words act in a similar manner to compounds. We included these as loan items in their own right, rather than splitting them into individual words. Hereafter, we use the term "loan" to refer to both individual loanwords and multi-word loan-phrases.

The code for extracting loan-phrase frequencies has some limitations, including the fact that English words are sometimes erroneously detected as Māori. This in turn affects which instances are counted (or not) in the co-occurrence analysis (Section 3.3). For instance, the script incorrectly extracted *hope more maori* as a nonce loan-phrase of size three, when the first two words were actually English (the wider context being “we *hope more Māori* groups will be able to use this space for events and functions”).

Since many loans were not productively used in the corpus, we removed all but 108 items that occurred at least five times (**step three**).

Next (**step four**), we manually inspected the remaining loans to remove false positives: (i) non-loans that share Māori phonotactics (e.g. *make*) and (ii) most proper nouns referring to personal names (e.g. *Hone Pene*) and places (e.g. *Rotorua*), unless they had a suitable native English alternative. The keyword *Matariki* “Maori New Year” was also removed because it was used to identify the newspaper articles in the first place. Our approach is largely an onomasiological one (Geeraerts, 2010), whereby loans are considered in relation to their native (receiver language) counterparts. However, eight of the proper nouns identified (excluding *Matariki*), do have counterparts available, and were therefore retained (e.g.

Aotearoa "New Zealand", *Māori* "native" and its highly productive, hybrid-derived counterpart *non-Māori*). Although our loans can largely be considered non-catachrestic (Onysko & Winter-Froemel, 2011) because they all have near-synonyms, not all loans have *perfectly* synonymous lexicalised (single-word) counterparts (e.g. *Pākehā* "New Zealand European"), nor are their English counterparts always productively used (e.g. *kawakawa* is seldom referred to as a "pepper tree").

Finally, we extracted plural forms (e.g. *Kiwis, maraes*) and combined loans with variant forms but the same meaning (e.g. *kaupapa* was merged with *kaupapa maori* to form *kaupapa (maori)* "Māori methodologies"). This resulted in a final list of 44 loans for consideration in our co-occurrence analysis (see Appendix 1 for details).

3.3 Computing Loan Co-occurrence

Next, we extracted patterns of co-occurrence by considering the newspaper articles in which the loans were used. To this end, we computed which loans from our list occurred in each text, ignoring directional relationships and searching the entire article, regardless of its size (see Section 1). In order to capture as many instances as possible, all text was lower-cased and macrons (indicating vowel length) were removed. This generated a co-occurrence matrix with 44 columns (loans) and 194 rows (texts). Although we calculated the exact frequency of each loan, this was ultimately treated as a binary interaction: the loan was either present in the text or not. In other words, it did not matter how many times a loan occurred in a particular text, as long as it appeared at least once.

Texts containing fewer than two loans from our list of 44 items were then excluded because they did not constitute a valid 'set' of loans. There were 18 articles (9.3%) that did not contain any loans apart from *Matariki* and 51 articles (26.3%) that contained only one loan in addition to *Matariki*. This left 125 articles (64.4%) for consideration in our co-occurrence analysis; see also Figure 5. Since only loans from our list of 44 items were considered, it is likely that even some discarded texts comprised two or more Māori loans (including at least one infrequent loan), but we wanted to focus on more general patterns of co-occurrence. Nevertheless, this does mean that the number of loans recorded per text is likely to be an underestimate of the true quantity.

The data from the resulting co-occurrence matrix was used to generate visualisations in the form of networks (Section 4.3) and hypergraphs (Section 4.4). For the networks, this

involved flattening each loan set into pairwise co-occurrences and calculating the frequency ('weight') of each pair. For the hypergraphs, we preserved the entire sets, and calculated their size and frequency.

3.4 Linguistic Properties

Following previous work, we identified three linguistic properties that are relevant to Māori loanword use in NZE (see Section 2.2), namely semantic domain, size and listedness. We coded our set of 44 loans with respect to each of these variables; see Appendix 1.

The first linguistic property coded was **semantic domain**. Macalister (2006) proposed four categories for typical Māori loans in NZE: flora and fauna terms (e.g. *kawakawa* "pepper tree"), proper noun terms (e.g. *Aotearoa* "New Zealand"), material culture terms (e.g. *taonga puoro* "musical instrument") and social culture terms (e.g. *kōhanga (reo)* "Māori immersion kindergarten"). The last two categories are not always straightforward to disambiguate; the crucial difference between them is that the former constitute a physical, concrete object that can be touched, whereas the latter do not. *Waka* is an interesting example because it traditionally refers to a wooden canoe, but can sometimes mean *any* form of transport, and, more recently, it has come to embody a general collective movement, as seen during the COVID-19 pandemic (Perkinson, 2020). Semantic changes of loans upon entering a receiver language have indeed been noted in previous work on NZE (Macalister, 2009; Author, 2020) and in other contact phenomena (e.g. Kurtböke & Potter, 2000).

Next, we coded the **size** of each loan by counting the number of words (following Author, 2017). This was straightforwardly applied based on spelling conventions: e.g. *iwi* "tribe" is size one; *kapa haka* "traditional Indigenous dance" is size two.

The final linguistic characteristic coded was **listedness**, following Muskyen (2000), and operationalised according to Stammers & Deuschar (2012). In our case, this is a binary variable denoting presence ('yes') or absence ('no') in *The New Zealand Oxford Dictionary* (Deverson & Kennedy, 2005).

Figure 2 shows the distribution of the 44 loan types when grouped by each of the three linguistic properties. Semantically, most loans are social culture terms (n=25, 57%), with the next most frequent categories, proper nouns and material culture loans, containing eight and seven loan types, respectively. The remaining four loans are flora and fauna terms. In terms of length, all but seven loans in our data are of size one (n=37, 84%), in keeping with typical

language contact phenomena observed elsewhere. Finally, the vast majority of loans (n=36, 82%) are listed in the dictionary.

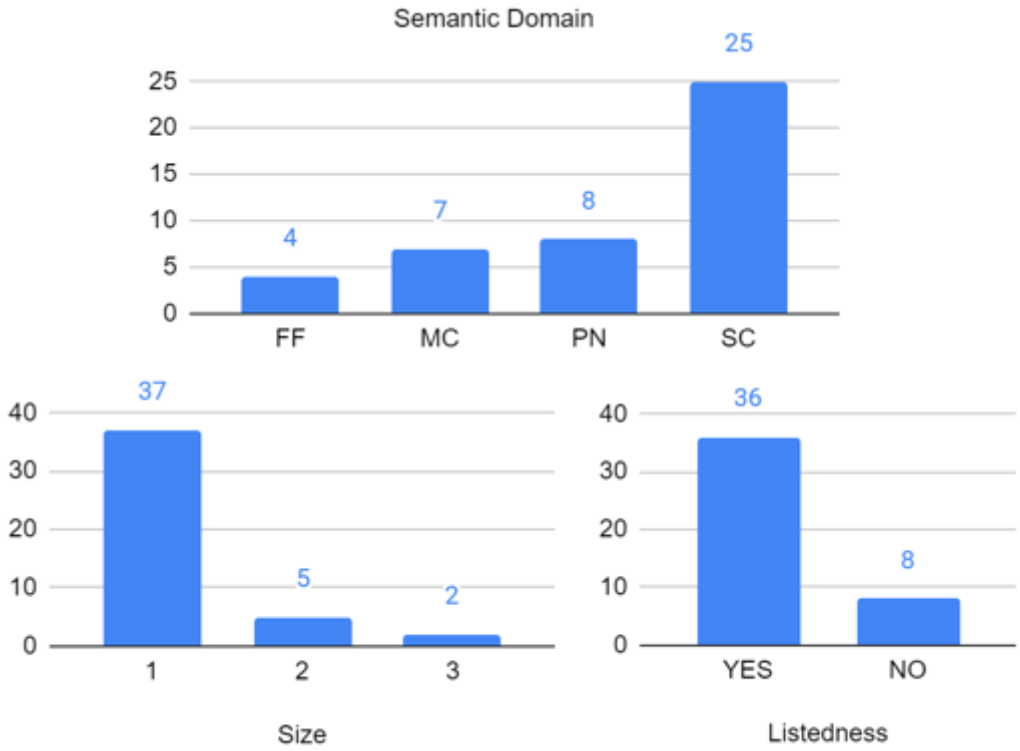

Figure 2. Linguistic properties of the 44 loan types of interest ('FF' = flora and fauna, 'MC' = material culture, 'PN' = proper noun and 'SC' = social culture)

We now contrast the number of types per category with the number of tokens per category, summarised in Figure 3. Since *Māori* is an outlier in the corpus (see Figure 4), we display counts both with and without this loan, using grey and blue bars, respectively. This distinction is important because we use the data represented by the grey bars for the network analysis (Section 4.2) and the data represented by the blue bars for the hypergraph analysis (Section 4.3). Unsurprisingly, given the dominance of single-word and listed loans, both size and listedness have similar distributions with respect to number of tokens. However, as regards semantic domain, it is proper nouns that are the most frequent when *Māori* is included, despite having relatively few loan types. Social culture terms still dominate when *Māori* is removed, however, and proper nouns then become the second least frequent category.

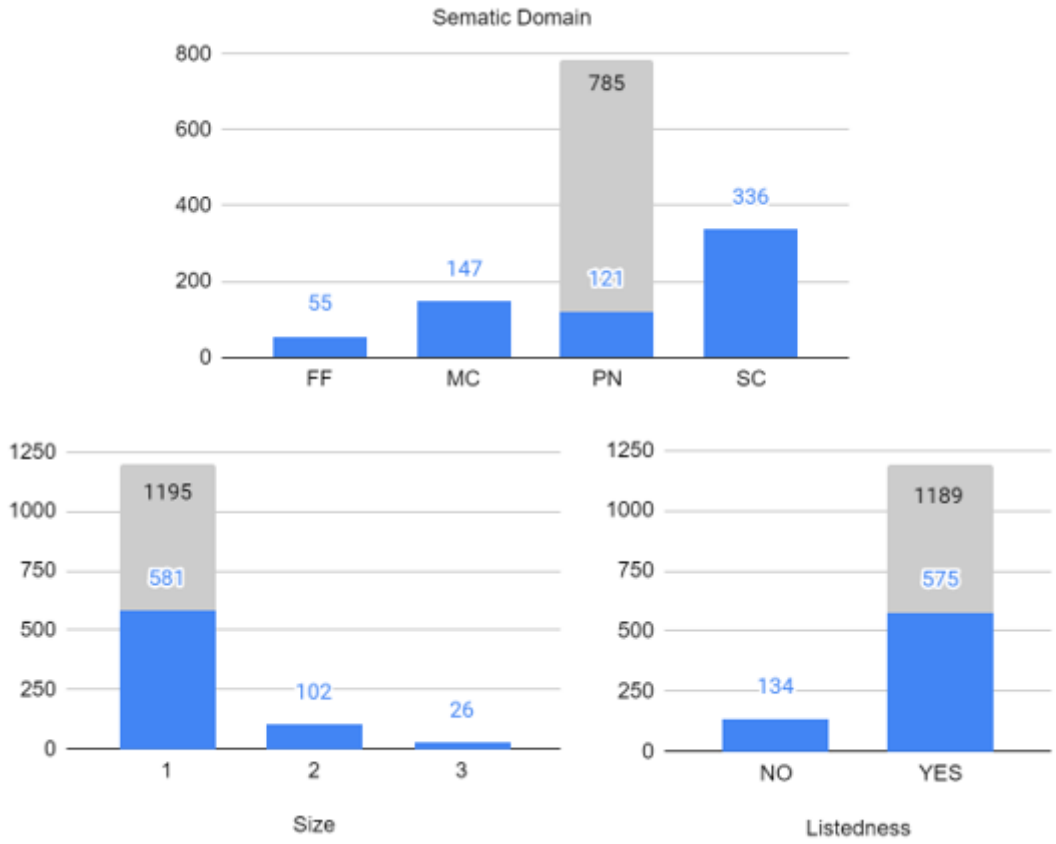

Figure 3. Linguistic properties aggregated by number of tokens per category

3.5 Overview of Loans by Frequency

Next, we summarise the overall frequency of the 44 loans in our list. There is a strong positive correlation between frequency and dispersion, such that frequent loans tend to occur in a greater number of texts than infrequent loans (Spearman R=0.77, t=8.03, df^4=41, p=6.037e-10). Collectively, the 44 loans occur 1,323 times in the corpus, with roughly two-thirds of loans occurring at least 10 times (including tokens arising from articles with only one loan). Figure 4 shows the raw frequency of all 44 productively-used loans in the Matariki Corpus. Of these loans, *Māori* "native" is by far the most frequent (n=614), followed by *Puanga* "Rigel Star" (n=50), whose rising is celebrated by some Māori tribes as an alternative to *Matariki*, and then *Kiwi* "New Zealand(er)" (n=44). The relatively high frequency of *Puanga*

is clearly linked to the topic of this corpus, and is much higher than we would expect to see in other contexts. Note that all loans in the figure have been lower-cased, including proper nouns, and macrons have been removed[5].

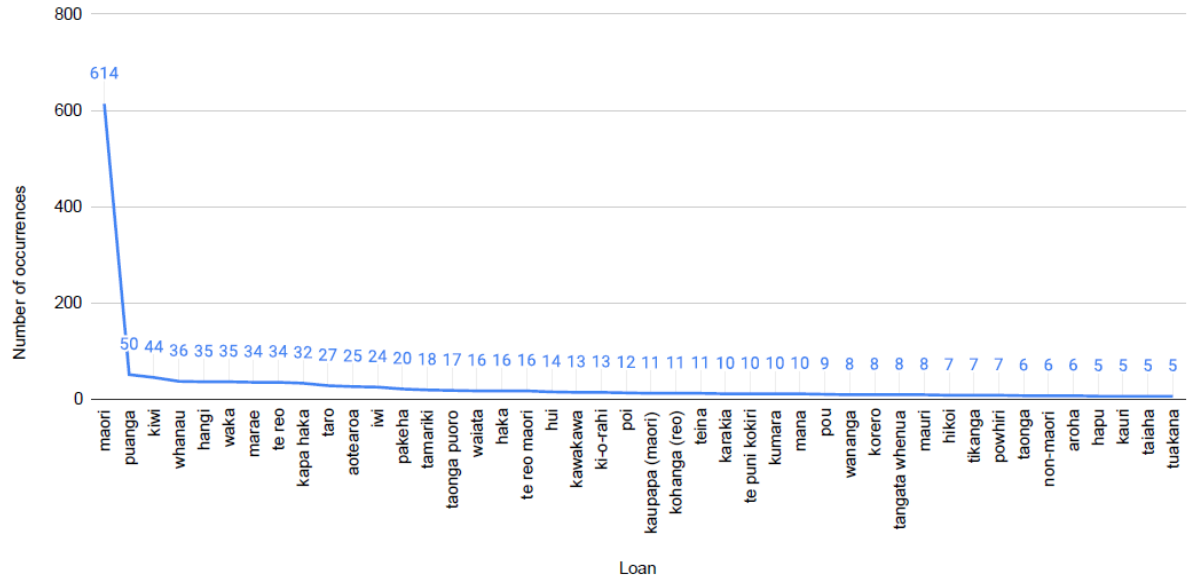

Figure 4. Raw frequency of productive loans in the corpus

4. Findings

In this section, we refer to the precise combination of loans in a text as a 'set'. We begin by summarising the number of loans per article (Section 4.1), then present standard networks, in which the loan sets are 'flattened' into pairwise co-occurrences (Section 4.2). Finally, we take a closer look at the loan sets as a whole, using a more robust representation, albeit less widely used in language analyses, called a 'hypergraph' (Section 4.3).

4.1 Distribution of Loan Types

We can study co-occurrence relationships by looking at the number of loans in each text. Figure 5 shows that, among the 194 newspaper articles and 44 loans, there are more articles that contain exactly one loan than any other number (roughly a quarter), followed by articles that contain two loans, and then three. The figures shown are conservative counts because only productively-used loans are considered. However, as seen in the right-hand

panel of Figure 5, most articles do contain at least two loans; this itself suggests that a network approach to studying loanwords may be appropriate. On average, each article contains 2.8 loans from our list of 44 items (with a median of two), or 3.9 loans if articles comprising fewer than two loans are ignored (with a median of three). The distribution of green bars on the left-hand panel of Figure 5 is right-skewed, showing that there is an inverse relationship between the number of loans in a text and the number of texts containing that many loans.

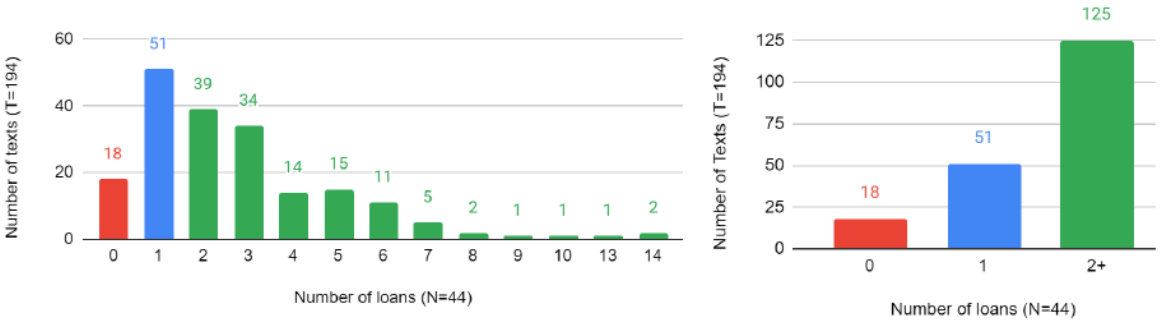

Figure 5. The number of loan types per text, including texts with no loans (red) and one loan (blue), which are omitted from our analysis

Moreover, if the data is examined from an individual loan perspective instead of a text perspective, we find that nearly 80% of the loan types in our list of 44 items (n=35, 79.5%) *never* occur in a text by themselves (e.g. *te reo* "language" does not appear in a text without at least one additional loan type), and all loans except *Māori* occur by themselves in fewer than four texts. *Māori* is present by itself in 38 texts, which accounts for roughly three-quarters of all texts containing a solitary loan (the blue bar in Figure 5). These statistics reinforce the observation that loans tend not to occur in isolation, highlighting the potential value of adopting a network approach to studying loanword use.

4.2 Standard Network Analysis: Pairwise Loan Co-occurrence

We now use standard networks to analyse patterns of pairwise loan co-occurrence in the Matariki Corpus. Classically, a network graph G is a pair $G = (V,E)$ where V is a set of 'vertices' and E is a set of 'edges' made up of pairs of vertices (see West, 1996). Thus, in a standard network, each edge connects exactly two nodes. We use the term 'node' to refer to vertices, and the term 'link' to refer to edges, as we believe these terms are more intuitive. Nodes can be thought of as entities of interest, and links as interactions between them. In our

case, nodes represent the 44 loans of interest and links represent the (bidirectional) co-occurrence of two loans within the same text (see Figures 6-8). Nodes closer to the centre of the network co-occur with a larger number of nodes than those at its periphery. The networks provide visual clues about the attractive force of the different loans, and the relationships between them.

There are several techniques for encoding additional information ('attributes') about the nodes and links in a network (see Nobre et al., 2019). For instance, we use node colour to denote one of our three linguistic properties, revealing how each category is distributed throughout the network. In addition, the thickness of each link is proportional to the number of texts featuring the corresponding pair of loans. This adds another layer of heterogeneity to the network than what can be observed from the topological effects alone. Finally, node size is proportional to loan frequency across the entire corpus, including tokens arising from texts containing only one loan.

All networks in this section were processed using the Python library *NetworkX* (Hagberg et al., 2008), and rendered using the open-source software package *Gephi* (Bastian et al., 2009). The final network layout implements the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), a force-directed technique, whereby nodes can be thought of as charged particles that repel one another, and links as springs that pull them together. Node positioning is continually refined until the system's overall energy (or 'stress') is minimised. This technique is non-deterministic, meaning that a different pass of the algorithm will yield a (slightly) different network. In practical terms, the configurations below cannot be reproduced exactly, but should nevertheless faithfully capture the networks' overall structure and complexity.

The most striking observation in Figures 6-8 is that all loans are connected, either directly or indirectly, such that each network consists of a single component. All three figures are identical apart from node colour, which is used to encode semantic domain, loan size and listedness, respectively. Predictably, *Māori* is at the centre of the network, and is, in fact, directly connected to every other loan. This means that all loans are at most two connections away from one another. Even if *Māori* were removed, there would still be one distinct cluster, but the distance between some nodes would increase to three connections. The strongest pairing is between *Māori* and *whānau* "family", which co-occur in 25 texts. In fact, of the 31 node pairs that occur in at least six texts, all but three involve *Māori*.

While the most frequent loans dominate the network, frequency is not always an indicator of node centrality. For instance, *iwi* "tribe" and *haka* "tribal war-dance" are relatively

infrequent yet very central, being connected to 35 and 25 loans, respectively. Conversely, despite having greater frequencies, *taro* "plant used for making bread" and *taonga puoro* "musical instrument" are more peripheral, being connected to 2 and 17 loans, respectively. This leads us to believe that 'degree' (the number of direct neighbours belonging to a given node) may provide a measure of entrenchment of different loans. The degree parameter (Appendix 2) can be thought of as a dispersion measure in regard to co-occurrence with other loans (rather than the number of texts a given loan occurs in).

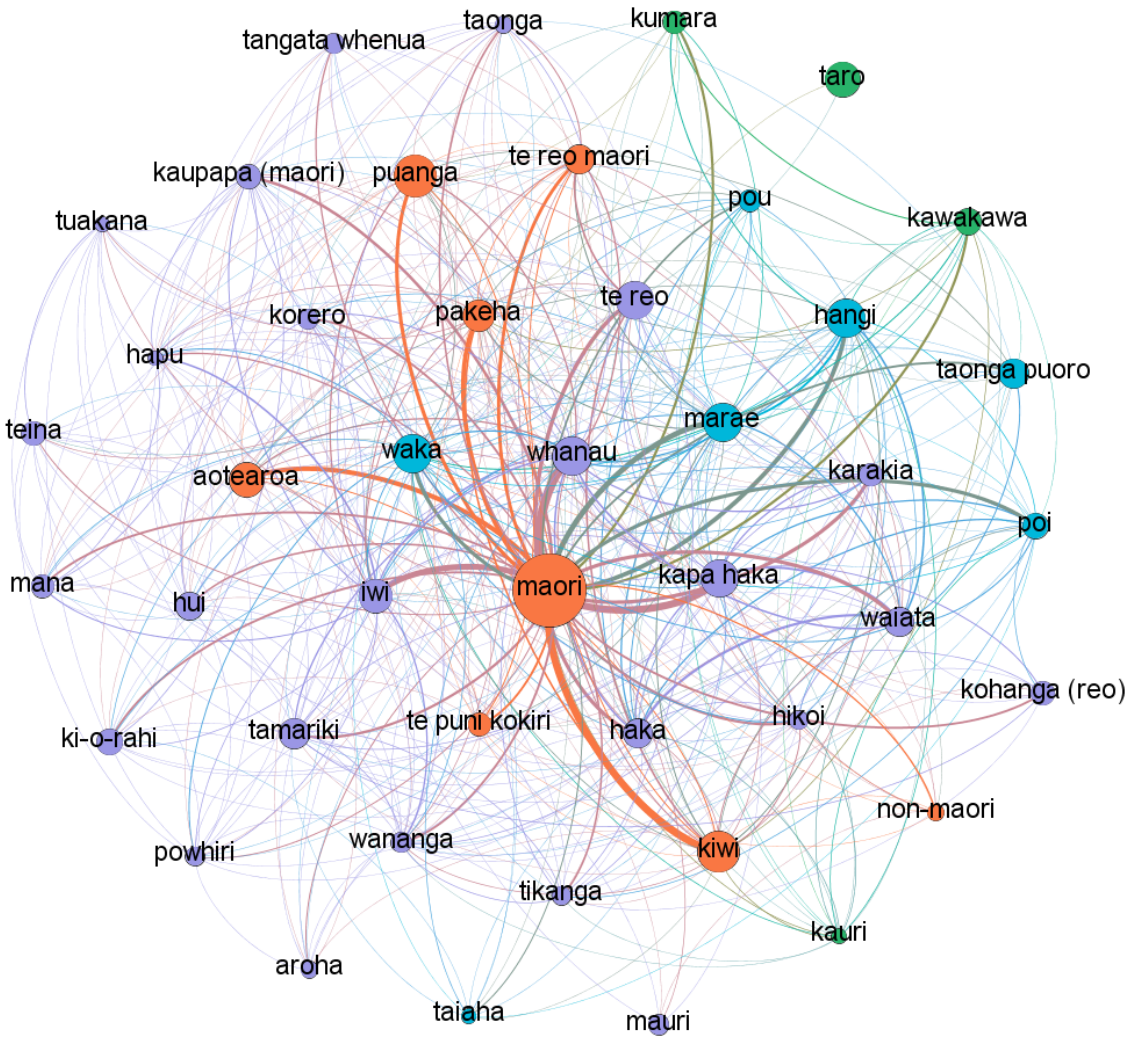

Figure 6. Standard network encoding semantic domain (social culture = lilac, proper noun = orange, flora and fauna = green, material culture = blue)

Looking at Figure 6, it is hard to determine whether nodes from the same semantic domain tend to be more strongly intra-connected. However, the only category that does not have

peripheral nodes is proper nouns, with the exception of the (less frequent) hybrid loan *non-Māori*. *Māori* is most strongly connected to the social culture loans *whānau* "extended family" (25 texts) and *kapa haka* "traditional Indigenous dance" (20 texts) as well as to *Kiwi* "New Zealand(er)" (20 texts), which is also a proper noun. Material culture and social culture loans occur in a mixture of positions (both central and peripheral), whereas flora and fauna terms are never central. The three strongest pairs that do not feature *Māori* occur between the social culture terms *haka* "war dance" and *waiata* "song" (7 texts), and *kapa haka* and *waiata* (6 texts), and the material culture terms *hāngī* "underground oven" and *marae* "meeting house" (5 texts).

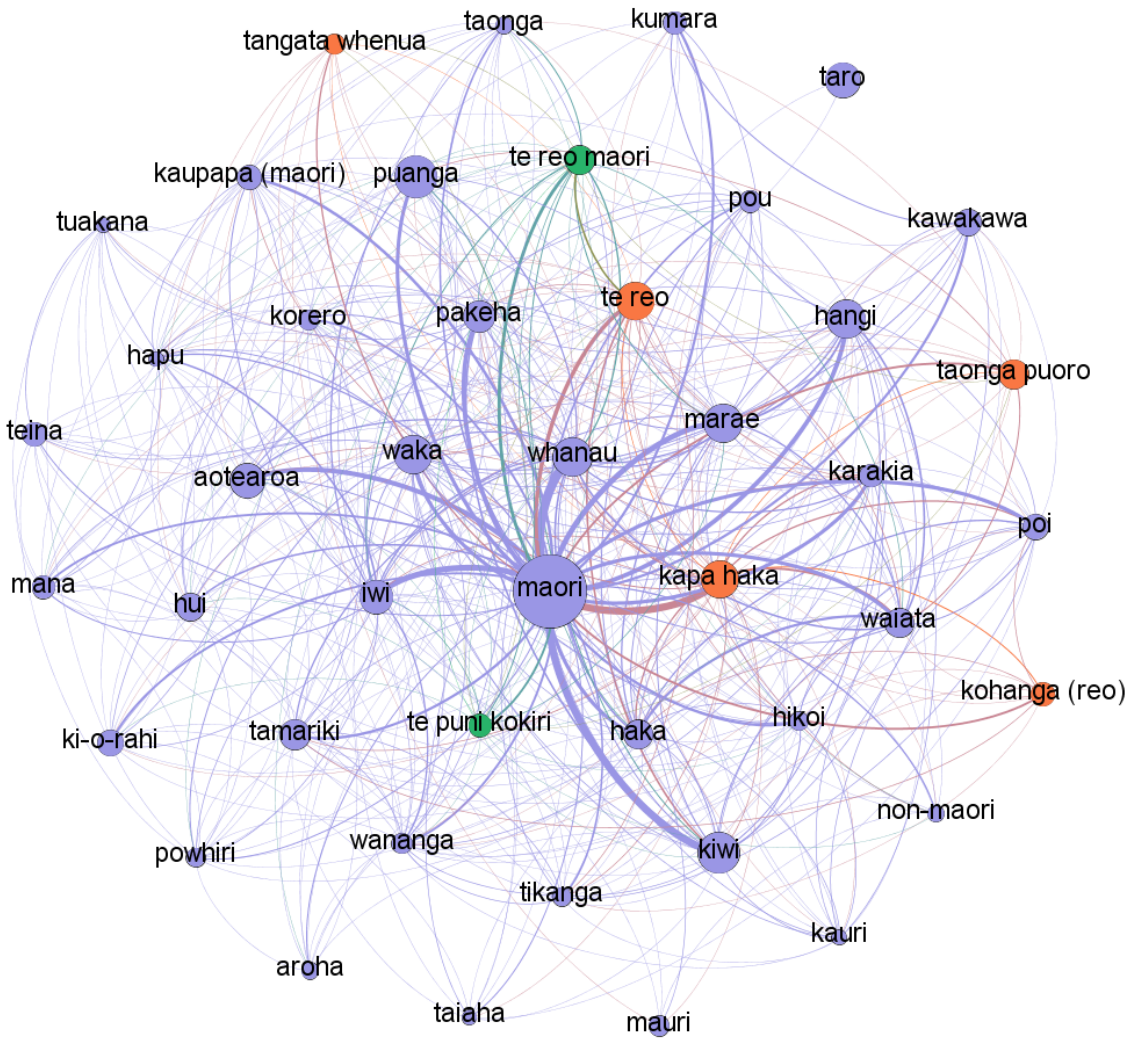

Figure 7. Standard network encoding loan size (one word = lilac, two words = orange, three words = green)

Figure 7 is heavily dominated by single-word loans, and, for the few loans that are made up of multiple words, it is harder to detect clustering patterns. We can also see that some loans comprising two or three words are frequent but not especially central (e.g. *te reo* "language", *te reo Māori* "the Māori language" and *taonga puoro* "musical instrument"). Conversely, *Te Puni Kokiri* "Ministry of Māori Development" is very central, despite being relatively infrequent.

Listedness in the dictionary confirms expected patterns (Figure 8), namely that listed (and more familiar) loanwords are generally more central in the network, and unlisted (less familiar, possibly newer borrowings) are more peripheral.

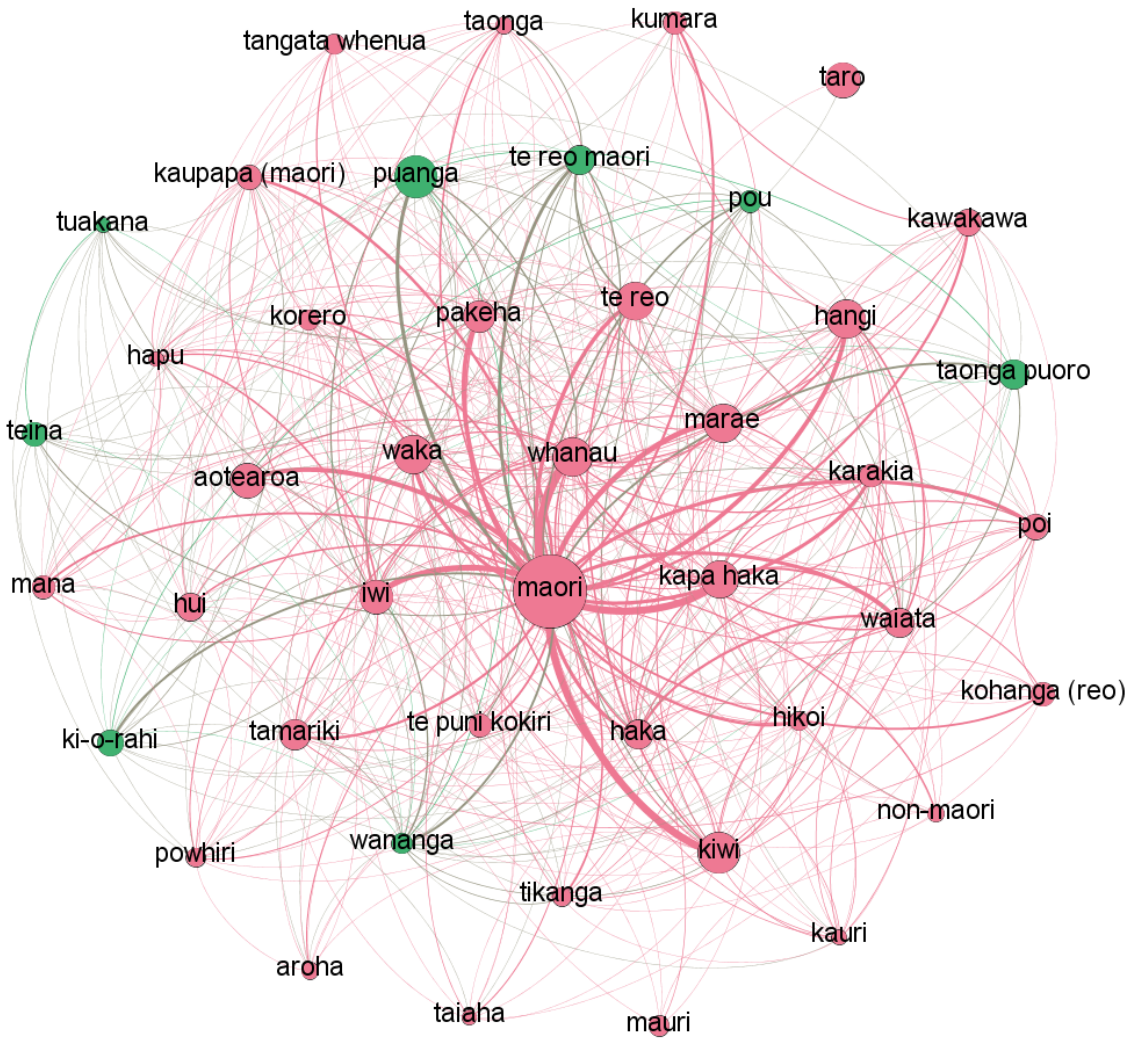

Figure 8. Standard network encoding listedness (listed = pink, unlisted = green)

Networks can also be explored from a statistical perspective, as given in Table 2. Apart from 'total edges', these metrics do not take into account the weight of each edge; the links connecting nodes are treated as binary interactions. 'Network density' captures how "tightly knit" the network is, expressed as a ratio of possible node pairings: i.e. 51% of all possible loan pairs are attested in one or more texts. The 'average degree' shows that each loanword, on average, is connected to 22 others (i.e. half of all loans), and this figure only decreases very slightly (to 20.5) when Māori is removed. Generally, these figures suggest the network is dense, meaning loans are highly connected.

Table 2. Network statistics for loans with at least five occurrences

Metric	**Value with *Māori* included (as per Figures 6-8)**	**Value with *Māori* removed**
Nodes	44	43
Distinct edges	483	440
Total edges	1,042	700
Average degree (range: 0-max nodes)	21.95	20.47
Network density (range: 0-1)	0.51	0.49
Triadic closure (range: 0-1)	0.65	0.64

4.3 Hypergraph Analysis: Preserving Sets of Loans

While networks tell us about the loans as individual items or as pairs, they inevitably result in information loss about the loans as a group (or 'set'). For instance, it is evident that *Māori* and *whānau* occur together in a large number of texts, but it is not clear if other loans are also present in those texts, and, if so, how many times each combination occurs. Figure 5 shows that roughly two-thirds of all sets contain more than two loans. Thus, a more faithful network representation would preserve information about the size and composition of these higher-order relationships.

To overcome this problem, we turn to the notion of a 'hypergraph'[6] (Berge, 1973). A hypergraph extends the above definition given for networks, allowing an edge (or 'hyperedge') to join multiple nodes, instead of just two. In mathematical terms, $G = (V,H)|H$,

where *H* is a set of *h* hyperedges comprising two or more vertices (as cited in Valdivia et al., 2019: 2).

To the best of our knowledge, hypergraphs have not previously been used in traditional linguistic analyses; however, they have been employed in computational studies (e.g. for modelling word-sense induction and other Natural Language Processing problems; see Qian et al., 2014; Soriano-Morales et al., 2014). These studies typically use hypergraphs to make predictions, but tend not to visualise them directly. We propose a novel application of hypergraphs, namely to visualise and analyse sets of loan co-occurrence. One software tool that can be used to this end is the online tool PAOHVis[7] (Valdivia et al., 2019), which can represent complex data sets involving up to 500 nodes. This is sufficient for visualising the 44 loans (nodes) and 125 sets (hyperedges) in our data. However, because *Māori* is so dominant, occurring in 117 of the 125 sets (93.6%; see Appendix 3), we have removed it from the following analysis, leaving 90 texts with two or more of the remaining 43 loans.

Figure 9 shows the total number of sets in which each loan occurs after removing *Māori*. The loan that occurs in the most sets is *whānau* (n=26, 28.9%), followed by *iwi* and *kapa haka* (n=16, 17.8%). Overall, comparing these values with Figure 4, it would appear that even infrequent loans are widely spread—relative to their frequency—among texts containing multiple loans.

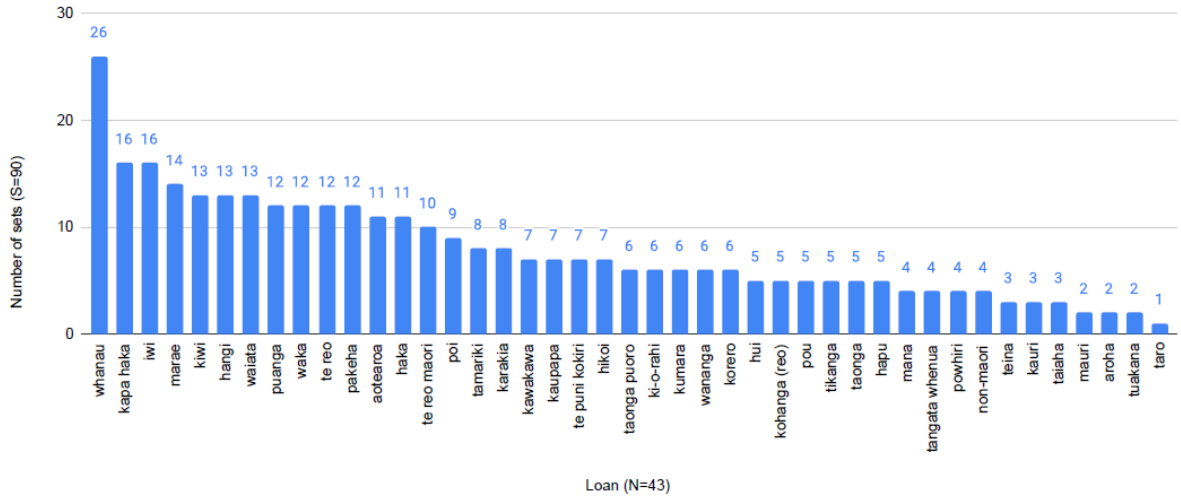

Figure 9. Loans by total number of sets (excluding the outlier *Māori*)

Unsurprisingly, given that we have already established a strong link between frequency and dispersion, and we know that loans tend not to occur in texts by themselves, there is a strong positive correlation between a loan's raw frequency and the number of sets in which it occurs (Spearman R=0.87, t=6.88, df=41, p<0.001). Analysis in PAOHVis reveals that over half of

all sets (n=49, 54.4%) contain at least one of the three most influential loans. Figures 10-11 show all 90 multi-loan sets in our data, coloured by semantic domain. Here, nodes (loans) are represented as parallel, horizontal bars, and hyperedges (loan sets) are denoted as vertical lines, with dots showing connections to one or more nodes. The number of loans in a set can be determined by counting the number of dots in a vertical line. Many sets contain loans from a mixture of semantic domains. In Figure 10, nodes are ranked according to the number of hyperedges (sets) they occur in; the numbers on the left-hand side therefore reflect the values shown in Figure 9. It is clear that the three most influential loans are all social culture terms (*whānau, kapa haka, iwi*), followed by two material culture terms (*marae, hāngī*), another social culture term (*waiata*), and a proper noun (*Kiwi*). Figure 11 shows the same data, but with the nodes arranged by importance in their categories, enabling comparisons within and between the various groups.

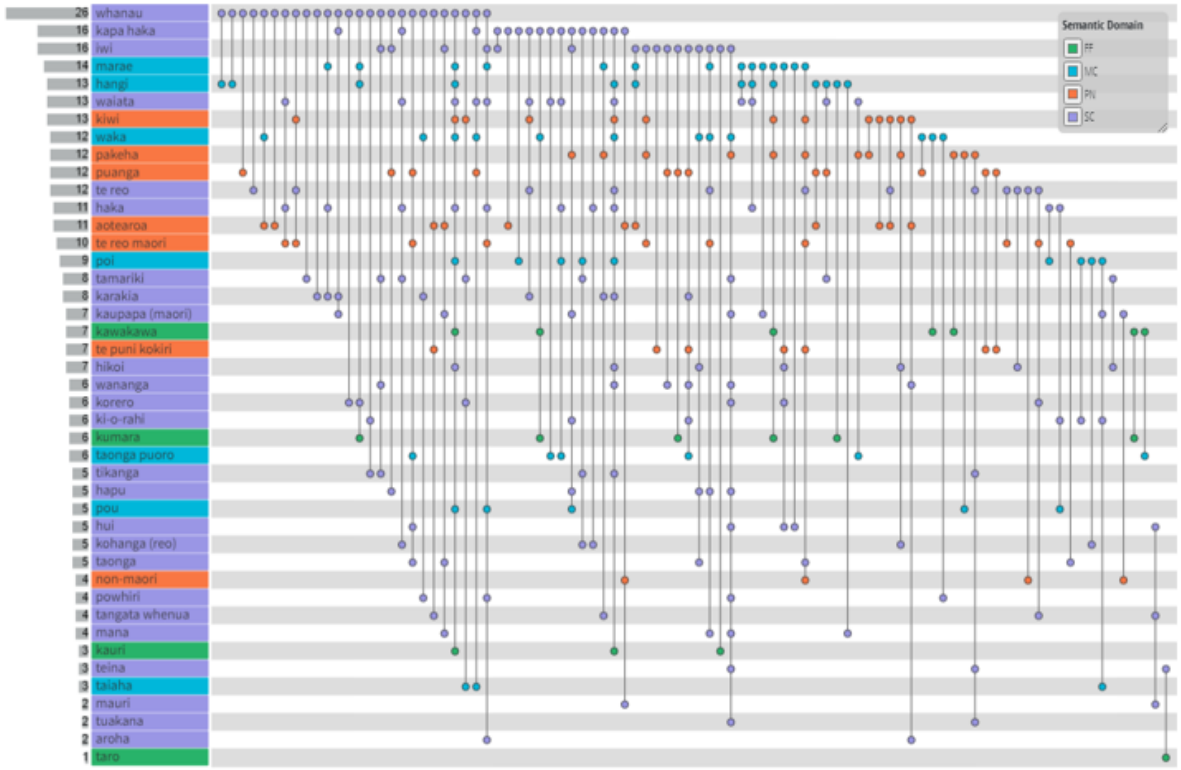

Figure 10. PAOHVis hypergraph showing all 90 sets coloured by semantic domain

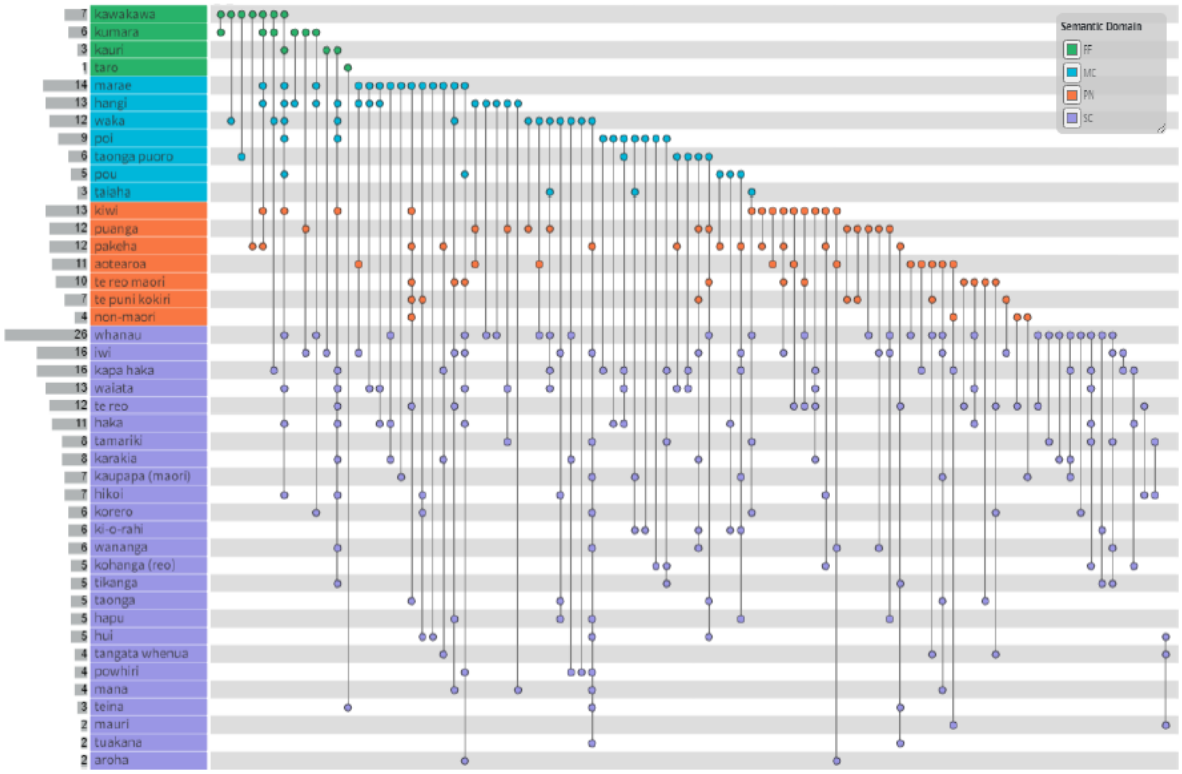

Figure 11. PAOHVis hypergraph, this time showing sets ordered by semantic domain

While we are constrained to using static figures in this paper, PAOHVis has several useful interactive features that facilitate exploration of the data. For instance, it is easy to reorder the data according to different metrics, to filter out less influential nodes, and to highlight all sets involving a particular node(s) of interest. Hovering over a node also reveals how many sets the corresponding loan has in common with every other loan. Input files and instructions for loading the loan sets into PAOHVis can be found at: https://github.com/Waikato/kiwiwords/tree/master/loanword_networks.

In addition to studying co-occurrence patterns among the 43 loans, hypergraphs can be used to investigate loan *categories*, by combining (or aggregating) nodes based on their linguistic properties. PAOHVis provides this functionality, but it does not support a simplified horizontal layout in which identical set configurations are consolidated into a single hyperedge. We therefore adapted the layout generated by PAOHVis, so that we could identify patterns about the various *kinds* of loans that tend to co-occur in our texts.

The remaining figures in this section are drawn by grouping the loans by their respective categories. Colour and shape are both used to emphasise the categories. As *Māori* is at the heart of most sets, occurring in all but seven, we removed it from the resulting hypergraphs to avoid undue influence from a single loan. When viewing these hypergraphs,

it is important to remember that the number of loans across categories is skewed, which means some categories are more likely to be present in a text (and to occur in greater numbers) than others. For instance, because there are six times as many social culture terms as there are flora and fauna terms, we would not expect to find many texts containing more flora and fauna loans than social culture loans.

Figure 12 shows a hypergraph in which loan sets are aggregated by semantic domain. In this representation, each coloured shape (node) indicates the presence of *any* loan type from the corresponding category, but it need not be the same loan across different texts. For each vertical line (set configuration), the number of instances of the same colour/shape tells us how many loans in the text belong to that category (e.g. a set comprising two orange circles has two proper nouns). The number of different colours/shapes in a set configuration then shows how many distinct categories there are (e.g. a set with two colours/shapes contains loans from two different semantic domains). The bar chart above each set indicates the frequency of that particular configuration. For example, the left-most line and bar chart in the figure show that there are 10 texts containing exactly one material culture term and one social culture term (which, again, may differ across texts: e.g. *marae* and *whānau* in one text, and *waka* and *kapa haka* in another) and no loans from any other semantic domain. The set configurations are ranked by frequency, as shown by the decreasing bar lengths. The figure only provides set configurations that occur at least twice, so as not to draw attention to infrequent combinations.

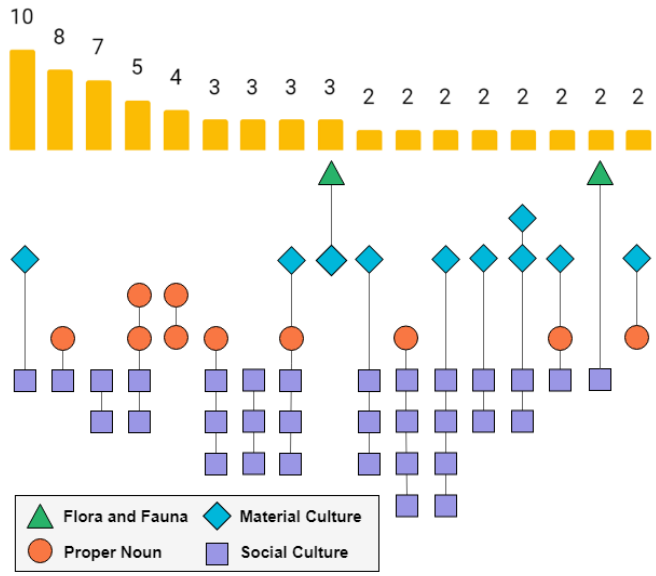

Figure 12. Loans aggregated by semantic domain (recurrent set configurations for 62 texts, 69% of the data)

Social culture loans are the most prevalent category in Figure 12, not only appearing in the largest number of set configurations, but also containing more occurrences within those sets (most commonly one or two occurrences, but sometimes up to four). This matches the high frequency of loan types and tokens seen in Section 3.4. In contrast to the general versatility and high presence of social culture loans, there is only ever one flora and fauna term in a text (two if we include the unique set configurations not pictured), and usually only one material culture loan, too; however, this is likely because there are fewer loans in these categories to begin with. No more than two proper nouns appear within a text, but recall the strict criteria for their selection. Most set configurations in Figure 12 involve loans from just two of the four categories, and none has loans from all four categories. However, Figure 12 only shows set configurations that are shared by two or more texts, which applies to just over two-thirds of the data (the yellow bars add up to 62, and therefore represent 69% of the 90 sets). Unsurprisingly, the remaining 28 set configurations—which all occur once—generally contain more loans (i.e. are larger sets), and have a higher maximum number of occurrences within each category (up to 11 social culture loans, five material culture loans, five proper nouns and two flora and fauna loans).

We can further aggregate this hypergraph by combining multiple loans from the same category into a single node (Figure 13). Here, nodes represent one *or more* loans from the corresponding category. This shows more general patterns, and includes data for all 90 sets. The most frequent combinations involve a mixture of social culture loans and proper nouns, and of material culture and social culture loans. The behaviour of material culture loans and proper nouns is quite similar with regard to the number of identical configurations they are part of. Flora and fauna terms are generally less dominant, the only category not to appear in any of the top five configurations. This is in line with the flora and fauna category comprising the smallest number of both types and tokens. There are only two texts containing loans from all four categories across the entire corpus. One striking observation is that, despite the high frequency of social culture loans in the data, most texts do not *only* contain social culture terms (only 13 from 90 do), but are instead accompanied by at least one loan from another semantic category. Unlike the other categories, there are no sets featuring *only* material culture loans, because there is no individual blue diamond in Figure 13.

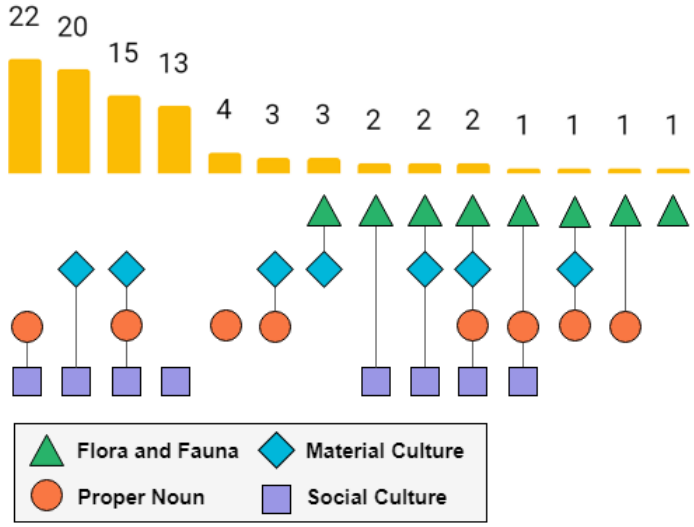

Figure 13. Loan sets collapsed by presence of semantic domain categories

The next set of visualisations (Figures 14-15) involve aggregating loans by size. Nearly all texts contain (one or several) loans of size one, with fewer texts containing loans of size two, and fewer still containing loans of size three. As a result, longer loans are always accompanied by shorter loans. Across the entire corpus (not shown in Figure 14), there are up to 13 loans of size one within the same text, but we never see more than two loans of size two or three.

Looking at Figure 15, there is a hierarchical structure, whereby sets are more frequent if they contain shorter loans from fewer categories. There are seven texts containing loans of all three sizes, and no texts containing only loans of size two or only loans of size three. This reinforces the pattern that phrases (i.e. loans of size two and three) do not appear in texts without there also being at least one single-word loan.

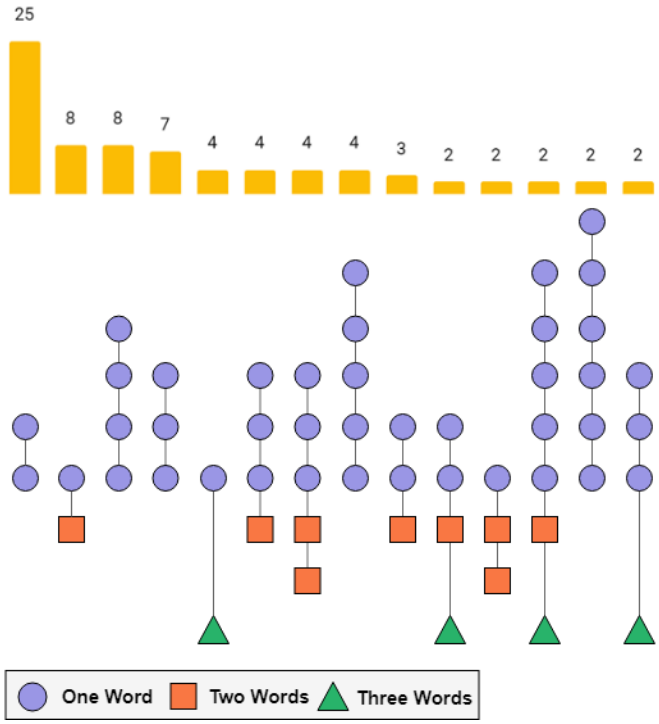

Figure 14. Loans aggregated by size (recurrent set configurations for 77 texts, 86% of the data)

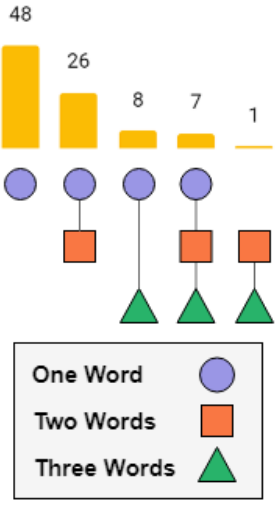

Figure 15. Loan sets collapsed by presence of size categories

Figures 16-17 show hypergraphs in which loans are aggregated by listedness in the dictionary. Recurrent set configurations vary between having only listed loans and a combination of one or two unlisted loans (Figure 16). Texts never contain *only* unlisted loans; they are always accompanied by one or more listed loans. While texts can have up to 12 listed loans (including unique configurations), the highest number of unlisted loans found within the same text is four. In fact, there are only three texts containing more unlisted loans than listed ones.

Although most sets solely consist of listed loans (58%), Figure 17 shows that there is still a large proportion of texts that contain a mixture of listed and unlisted items (42%).

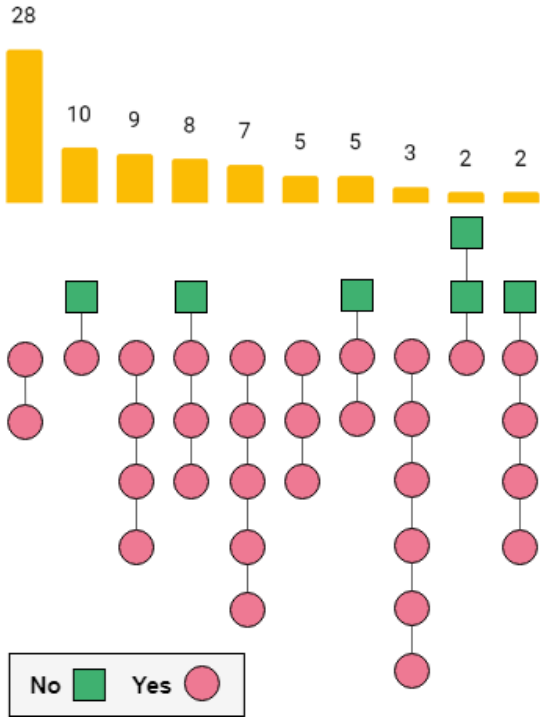

Figure 16. Loans aggregated by listedness (recurrent set configurations for 79 texts, 88% of the data)

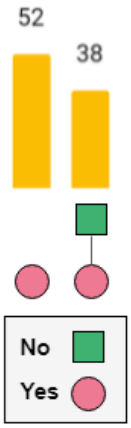

Figure 17. Loan sets collapsed by presence of listedness categories

In order to investigate patterns concerning the frequency profile of the loans in each set, the 43 loans of interest were divided into four different frequency bands, each containing 10-12 items, based on the loans' raw frequency rankings (see Appendix 1). Band 1 contains the most frequent loans and Band 4 the least frequent loans. According to Figure 18, none of the recurrent sets have more than three loans from any frequency band, which is perhaps unexpected for higher frequency bands. However, as before, this first hypergraph only considers recurrent set configurations; there are several much larger sets that occur only once.

Among the 36 texts with unique configurations (not shown), there are up to five loans in a single text from Band 1, four from Band 2, five from Band 3 and three from Band 4. Overall, each text appears to contain more loans from higher frequency bands.

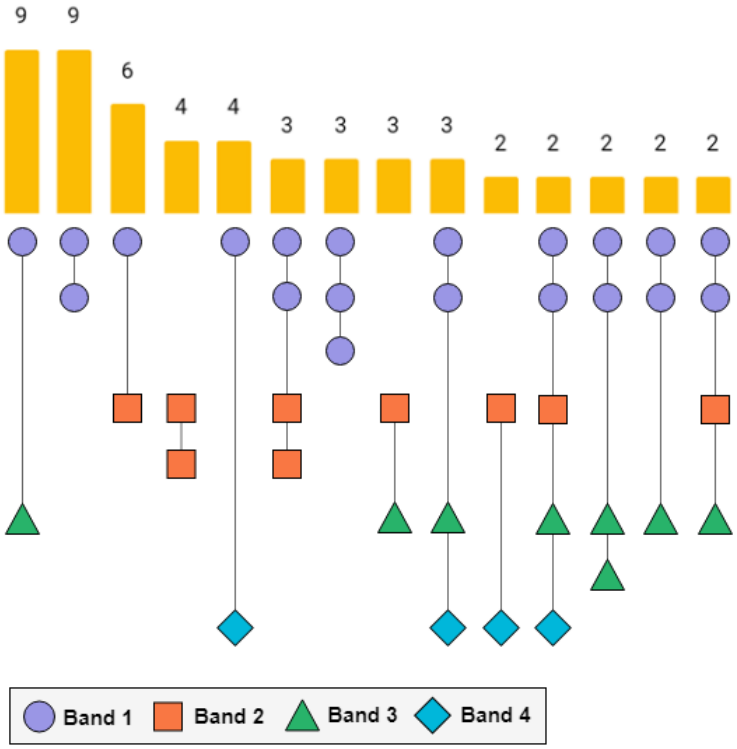

Figure 18. Loans aggregated by frequency bands (recurrent set configurations for 54 texts, 60% of the data)

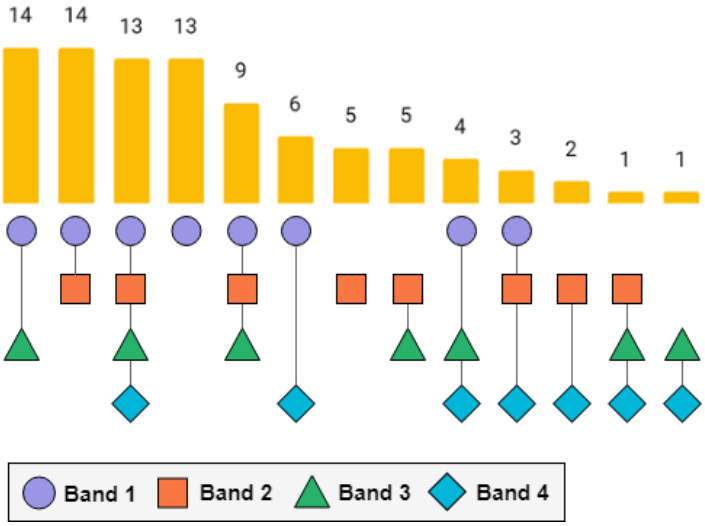

Figure 19. Loan sets collapsed by presence of frequency bands

Figure 19 considers only the presence or absence of each frequency band within a text. All texts except one contain loans from either or both of the first two frequency bands. Consequently, less frequent loans (Bands 3 and 4) nearly always occur in sets with one or more frequent loans (Bands 1 and 2), but not the other way round. Loans from Band 1 dominate, occuring in all of the six most common set configurations (more than three-quarters of the 90 sets). Loans in Band 4 are the least dominant, and nearly always occur with a loan from Band 1 (with only four exceptions, which still contain loans from higher frequency bands). Thus, infrequent loanwords tend to be accompanied by one or more frequent loanwords.

5. Discussion

This section begins with a summary of the main contributions of this study, focusing on the language contact situation in New Zealand, before turning to wider implications.

In Section 3, we investigated three linguistic properties that are relevant to the language contact situation in Aotearoa, namely semantic domain, size and listedness. Grouping the 44 loans of interest by each of these properties confirms trends observed in previous work, with the majority of productive Māori loan types being social culture terms, consisting of a single word, and being listed in the dictionary. In the latter case, these listed loans are unlikely to be new borrowings, having been in the language long enough to be included in dictionaries. One caveat here is that loans were only included if they had an existing English near-synonym; as such, most proper nouns were omitted from the analysis.

In general, looking at the number of loans per text, our data shows that most texts contain at least two loans, and that most loans do not occur by themselves in any articles. The median of two/three loans per text is almost certainly a conservative measure because of our strict criteria for loanword inclusion. The trends identified in the loanword co-occurrence networks reveal that all loans are connected, either directly or indirectly, and no loans are further than three connections away from each other. These networks provide a snapshot of the links between loans by ‘flattening’ the data into pairwise relationships, highlighting the centrality of certain loans and their categories.

Given that writers have an optional (native) English counterpart to the loans used in these texts, there is no *a priori* reason for the use of one loan to automatically draw out the use of another, as far as expression of meaning is concerned. Our intuition is that Māori loans

operate within a linguistic ecosystem in NZE, whereby speakers (or, in our case, writers) do not appear to make individual word choices, but rather adopt loans as a set.

Our findings fit proposals that argue against analyses of loans as "single lexemes" (e.g. Kurtböke & Potter, 2000: 88), which ignore the larger picture and consequences of relationships that hold between words. Although Kurtböke & Potter (ibid) limit their scope to collocates, looking to the immediate left or right of a given loan, their general point of adopting Sinclair's (1996) proposal for (monolingual) word use and extending it to language-mixing resonates in our data, too. It is also our view that taking loan use to represent a "slot-and-filler" basis misses co-textual links, be they links to immediate collocates (at a micro-discourse level) or links across a larger textual distance (at a macro-discourse level).

As regards the contact situation in New Zealand English, we interpret the motivation for the use of loans to be more aligned with ideology than Māori/English bilingualism (which is currently at 5%, according to Stats NZ, 2018). We propose that the choice of whether to use a particular loan or not is *not* made at the individual lexical level, but rather more globally, at the text level. This observation is also in line with Hashimoto's (2019) findings that speakers who use words of Māori origin and who attempt to pronounce them as they would be pronounced in Māori show greater affinity towards Māori language, culture and general worldview. While previous accounts of loanword use (Macalister, 2007: 504) suggest that speakers use loanwords for brevity, clarity, expression of identity, empathy or cultural reference, our work suggests that there may be an additional ideological factor in play: namely, overt alignment with Māori language and culture. Because the use of Māori loanwords is salient in the discourse, it is also a socially meaningful act and, as such, the presence of multiple loans within a text (rather than just one) serves to further highlight the ideology which accompanies such use. In particular, with an observed increase in loanword use in current times (e.g. Author, 2019), the motivations for these lexical choices are likely shifting.

Another observation is that the loans in the networks exhibit clustering with respect to some linguistic properties but not others. Arguably, the most surprising clustering concerns semantic domain: the fact that flora and fauna loans are never central, whereas social culture and material culture terms occur in a variety of positions in the network, while still being reasonably well-connected among themselves.

Furthermore, we find patterns of co-occurrence in the networks that are not predictable from overall frequency. Some loans are central despite being relatively infrequent (e.g. *iwi* "tribe", *haka* "tribal war-dance"); for others, the opposite is true (e.g. *taro* "plant used for

making bread", *taonga puoro* "musical instrument"). There is only one pathway for a loan to be central: namely, to co-occur with many other loans. However, external loans may be peripheral for one of several reasons, including the fact that they are incoming new loans (evidenced by their unlisted status) or that they entered NZE some time ago and occur only in niche textual environments. The flora and fauna term *kauri* is an example of the latter, being both entrenched and less intrinsically linked to Māori culture than many of the other terms, whereas *Puanga* is an example of the former.

Hypergraphs were employed to extend the capabilities of our networks, by preserving the entire set of loans in each text. This also meant we could aggregate the data in ways that networks do not allow, revealing the following overall trends:

i. Social culture terms dominate the loanword sets (as well as overall types and tokens), and often occur in texts with a material culture term or proper noun;
ii. Loan phrases are accompanied in a text by one or more single-word loans;
iii. While most texts contain only listed loans, over 40% of sets contain a mixture of listed and unlisted loans;
iv. Infrequent loanwords tend to be accompanied by one or more frequent loanwords (not least because most sets contain at least one loan from the highest frequency band).

With regard to finding (iii), the presence of unlisted loans indicates that we are still riding a wave of borrowing importation from Māori, and could possibly be at the beginning of a third wave, following on from the two initial waves proposed by Macalister (2006). The separation of borrowing waves is evidently not something that can be identified while the change is taking place; rather, it will only be diachronically that this shift may reveal itself. Further studies of loanwords in this language contact context will be needed to track the potential presence of a third wave. Finding (iv) suggests that loanwords occur in vocabulary frequency bands, not dissimilar from those proposed for measuring L2 vocabulary (see Laufer & Nation, 1995). This has implications for gauging loanword familiarity (Macalister, 2009), as knowledge of loans in a lower frequency band implies knowledge of loans in higher frequency bands.

Our analysis suffers from three main limitations. First, the corpus was obtained by tracking newspaper articles pertaining to Matariki, and this topic may have introduced certain biases in the loanword selection. For example, *Puanga* "Rigel star" undoubtedly has a much

higher normalised frequency in this corpus than it would in a different corpus. Conversely, the topic of Matariki may not lend itself to a high number of flora and fauna terms, as these are less intrinsically linked to Māori culture. A second, related limitation is that the number of loans in each category is highly skewed across the four linguistic properties coded (e.g. there are many more social culture terms than flora and fauna terms), which means our results favour the more well-represented categories. This problem is partly due to our loanword selection criteria (see Section 3.2), which aimed to make the data more manageable, while helping to investigate lexical choice. Nevertheless, these decisions came at the expense of reducing the size of our sets and excluding potentially relevant loans from the analysis. A final limitation is the fact that our approach treats all texts as though they provide equal opportunities for loan use, even though the articles differ in length: shorter texts provide fewer opportunities for loan use compared to longer texts. We did not make adjustments based on word counts, because we wanted to package each text as a whole without distorting the loan sets they contain in any way.

6. Conclusions

This paper introduced a novel methodological approach to studying loanwords. In order to test wider discourse-level patterns among Māori loanwords in NZE, we created network and hypergraph visualisations that explore patterns of loanword co-occurrence at the text-level. Our analysis has shown how networks and hypergraphs can be used to uncover fresh insights into loanword use, especially when sets and pairs of loans are analysed in relation to other linguistic properties, such as semantic domain and listedness. We believe that our findings complement traditional, frequency-based approaches, helping to shed light on hidden and complex patterns in a corpus by examining the data through a different lens.

Standard networks are useful for understanding how different entities interact, and they also provide a mechanism for encoding multiple attributes simultaneously. In our case-study, we used networks to show not only loan co-occurrence (via weighted links), but also overall frequency of use (by varying node size). Because humans are better at perceiving visual patterns than interpreting large tables of raw data—especially multi-dimensional data—networks of loan co-occurrence constitute a more insightful means of representing the underlying patterns.

One aspect of loanword co-occurrence that is not faithfully represented by standard networks is group-level phenomena (i.e. interactions between multiple nodes). This limitation also means that networks cannot be used to study the size and frequency of complete sets in relation to linguistic properties. Hypergraphs constitute an elegant solution for such an analysis, and we believe they also lend themselves to studying other linguistic phenomena.

These methods enable several opportunities for future work. One such avenue is diachronic analysis to determine, among other things, which (types of) loan sets are the most stable across time. This could be achieved through the application of 'dynamic hypergraphs' (Valdivia et al., 2019), although such an analysis may pose data sparsity issues. For the New Zealand context, it would be especially beneficial to examine more recent data than that captured by the Matariki Corpus, as our intuition is that Aotearoa is currently experiencing an attitudinal shift towards increased acceptance of the Māori language. A second avenue to consider is the potential significance of the *position* of loans in each text, and specifically, the extent to which the first loan used in a text may "trigger" the subsequent use of others, which could also be explored diachronically. There are opportunities for macro-discourse approaches to be used more widely in loanword studies, probing different genres and language pairs. Finally, as mentioned above, we believe networks and hypergraphs can be leveraged in other linguistic studies, both within and outside the area of loanword research.

Notes

1. *Aotearoa* is commonly used as the Māori name for New Zealand.

2. See https://github.com/TeHikuMedia/nga-kupu

3. The Māori alphabet consists of ten consonants (*h, k, m, n, ng, p, r, t, w, wh*) and five vowels (*a, e, i, o, u*).

4. The degrees of freedom used reflects the removal of the loan *Māori* from the data, which is an outlier (see Figure 4).

5. This is also the case for Figures 6-11.

6. A hypergraph is also called a 'family of sets' obtained from the universal set.

7. PAOHVis is accessible from http://paovis.ddns.net/paoh.html

References

Author (2017)
Author (2019)
Author (2019)
Author (2020)
Author (2020)
Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. In *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 3, No. 1).
Berge, C. (1973). *Graphs and Hypergraphs*, North-Holland, Amsterdam.
Chesley, P., & Baayen, H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics,* 48, 1343–1374.
Davies, C., & Maclagan, M. (2006). Maori Words - Read all about it: testing the presence of 13 Maori words in 4 New Zealand newspapers from 1997 to 2004. *Te Reo*, 49, 73–99.
de Bres, J. (2006). Maori lexical items in the mainstream television news in New Zealand. *New Zealand English Journal*, 20, 17–34.
Degani, M. (2010). The Pakeha myth of one New Zealand /Aotearoa: An exploration in the use of Maori loanwords in New Zealand English. In R. Facchinetti, David Crystal & Barbara Seidlhofer (eds.), *From international to local English – and back again*, 165–196. Frankfurt am Main: Peter Lang.
Denis, D., & D'Arcy, A. (2018). Settler colonial Englishes are distinct from postcolonial Englishes. *American Speech*, 93(1), 1–31. doi 10.1215/00031283-6904065
Deverson, T., & Kennedy, G. (2005). *The New Zealand Oxford Dictionary*. Oxford: Oxford University Press.
Firth, J. R. (1957). *Papers in Linguistics.* London: Oxford University Press.
Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*(11), 1129-1164.
Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford: Oxford University Press.
Görlach, M. (2002). *English in Europe.* Oxford: Oxford University Press.
Gries, S. (2013). 50-something years of work on collocations: What is or should be next…. *International Journal of Corpus Linguistics*, *18*(1), 137–166.
Gries, S. (2021). A new approach to (key) keywords analysis: using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2), 1–33.
Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX*.
Hashimoto, D. (2019). *Loanword phonology in New Zealand English: Exemplar activation and message predictability*. PhD Dissertation. University of Canterbury, NZ.
Haugen, E. (1950). The Analysis of Linguistic Borrowing. *Language,* 26(2), 210–231.
Kennedy, G. (2001). Lexical borrowing from Maori in New Zealand English. In Bruce Moore (ed), *Who's centric now? The present state of Post-colonial Englishes*, 59–81. Melbourne: Oxford University Press.
Kurtböke, P., & Potter, L. (2000). Co-occurrence tendencies of loanwords in corpora. *International Journal of Corpus Linguistics*, 5(1), 83–100.
Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.
Macalister, J. (2000). The Changing Use of Maori Words in New Zealand English. *New Zealand English Journal*, 14, 41–47.

Macalister, J. (2006). The Maori presence in the New Zealand English lexicon, 1850–2000: Evidence from a corpus-based study. *English World-Wide*, 27, 1–24.
Macalister, J. (2007). *Weka* or *woodhen*? Nativization through lexical choice in New Zealand English. *World Englishes*, 26(4), 492-506.
Macalister, J. (2009). Investigating the changing use of te reo. *NZ Words*, 13, 3–4.
MacDonald, D. E., & Daly, N. (2013). Kiwi, kapai, and kuia: Māori loanwords in New Zealand English children's picture books published between 1995 and 2005.
Muysken, P., & Muysken, P. C. (2000). *Bilingual speech: A typology of code-mixing*. Cambridge University Press.
Nobre, C., Meyer, M., Streit, M., & Lex, A. (2019). The state of the art in visualizing multivariate networks. In *Computer Graphics Forum* (Vol. 38, No. 3, pp. 807-832).
Onysko, A., & Winter-Froemel, E. (2011). Necessary loans–luxury loans? Exploring the pragmatic dimension of borrowing. *Journal of pragmatics*, 43(6), 1550–1567.
Perkinson, E. (2020). He waka eke noa! *Aotearoa New Zealand Social Work*, *32*(2), 71-72.
Poplack, S. (2018). *Borrowing: loanwords in the speech community and in the grammar.* Oxford: Oxford University Press.
Qian, T., Ji, D., Zhang, M., Teng, C., & Xia, C. (2014). Word sense induction using lexical chain based hypergraph model. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 1601-1611).
Sinclair, J. (1996). The search for units of meaning. *Textus* 9, 75–105.
Soriano-Morales, E. P., Ah-Pine, J., & Loudcher, S. (2016). Using a Heterogeneous Linguistic Network for Word Sense Induction and Disambiguation. *Computación y Sistemas*, *20*(3), 315-325.
Stammers, J. R. & Deuchar, M. (2012). Testing the nonce borrowing hypothesis: Counter-evidence from English-origin verbs in Welsh. *Bilingualism:Language and Cognition,* 15(3), 630–643.
Statistics NZ (2018). Profile of New Zealand 2018 Census – Māori Statistics. https://www.stats.govt.nz/2018-census/ [Accessed 24 June 2021]
Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N., & Fekete, J. D. (2019). Analyzing Dynamic Hypergraphs with Parallel Aggregated Ordered Hypergraph Visualization. *IEEE Transactions on Visualization and Computer Graphics*.
Weinreich, U. (1953) *Languages in Contact*. Mouton: The Hague.
West, D. B. (1996). *Introduction to graph theory* (Vol. 2). Upper Saddle River, NJ: Prentice hall.
Zenner, E., Speelman, D., & Geeraerts, D. (2012). Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of Anglicisms in Dutch. *Cognitive Linguistics*, 23(4), 749–792.
Zenner, E., Speelman, D. & Geeraerts, D. (2013). What makes a catchphrase catchy? Possible determinants in the borrowability of English catchphrases in Dutch. In Eline Zenner & Gitte Kristiansen (eds.), *New perspectives on lexical borrowing*, 41–64. Berlin, New York: De Gruyter.
Zenner, E., Speelman, D. & Geeraerts, D. (2015). A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show. *International Journal of Bilingualism*, 19(3), 333–346.
Zipf, George K. (1935). *The psychobiology of language.* Oxford: Houghton-Mifflin.

Appendix 1: Productive Māori Loans

The following 44 loans occur at least five times in the Matariki Corpus:

Loan	English Counterpart	Semantic Domain	Size	Listed	Frequency Band
Aotearoa	New Zealand	PN	1	YES	1
aroha	love	SC	1	YES	4
haka	war dance, tribal dance	SC	1	YES	2
hāngī	underground oven	MC	1	YES	1
hapū	sub-tribe, clan	SC	1	YES	4
hīkoi	walk, protests	SC	1	YES	4
hui	meeting	SC	1	YES	2
iwi	tribe	SC	1	YES	1
kapa haka	traditional Indigenous dance	SC	2	YES	1
karakia	prayer	SC	1	YES	3
kaupapa (Māori)	Māori methodologies	SC	1	YES	3
kauri	largest tree found in the North Island	FF	1	YES	4
kawakawa	pepper tree	FF	1	YES	2
kī-o-rahi	traditional game	SC	1	NO	2
Kiwi	New Zealand(er), pertaining to NZ, also the name of a flightless bird	PN	1	YES	1
kōhanga (reo)	Māori immersion kindergarten (lit. "language nest")	SC	2	YES	3
kōrero	talk, conversation	SC	1	YES	3
kūmara	sweet potato	FF	1	YES	3
mana	power	SC	1	YES	3
*Māori**	native, Indigenous	PN	1	YES	1
marae	meeting house	MC	1	YES	1
mauri	life force	SC	1	YES	3
non-Māori	non-Indigenous (esp. Pākehā)	PN	1	YES	4
Pākehā	New Zealand European	PN	1	YES	2

poi	ball on a string featured in musical performances	MC	1	YES	2
pou	support poles	MC	1	NO	3
pōwhiri	welcoming ceremony	SC	1	YES	4
Puanga	Rigel star	PN	1	NO	1
taiaha	long wooden weapon	MC	1	YES	4
tamariki	children	SC	1	YES	2
tangata whenua	people of the land	SC	2	YES	3
taonga	treasure	SC	1	YES	4
taonga puoro	musical instrument	MC	2	NO	2
taro	plant used for making bread	FF	1	YES	1
Te Puni Kōkiri	Ministry of Māori Development	PN	3	YES	3
te reo	language, voice	SC	2	YES	1
te reo Māori	the Māori language	PN	3	NO	2
teina	younger brother/sister (of same gender)	SC	1	NO	3
tikanga	custom	SC	1	YES	4
tuakana	elder brother/sister (of same gender)	SC	1	NO	4
waiata	song	SC	1	YES	2
waka	canoe	MC	1	YES	1
wānanga	university, learning seminar/conference	SC	1	NO	3
whānau	extended family	SC	1	YES	1

*****_Māori_ is an outlier in the corpus, and has been removed from parts of the analysis

Appendix 2: Loans by Degree

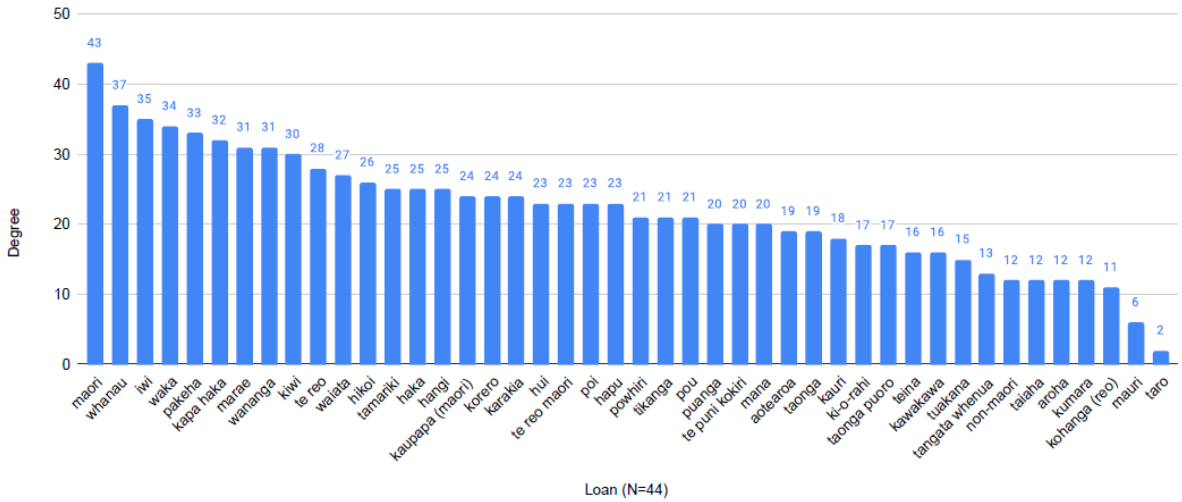

Figure 20. Productive loans in the corpus, ordered by number of connected nodes

Appendix 3: Sets including Māori

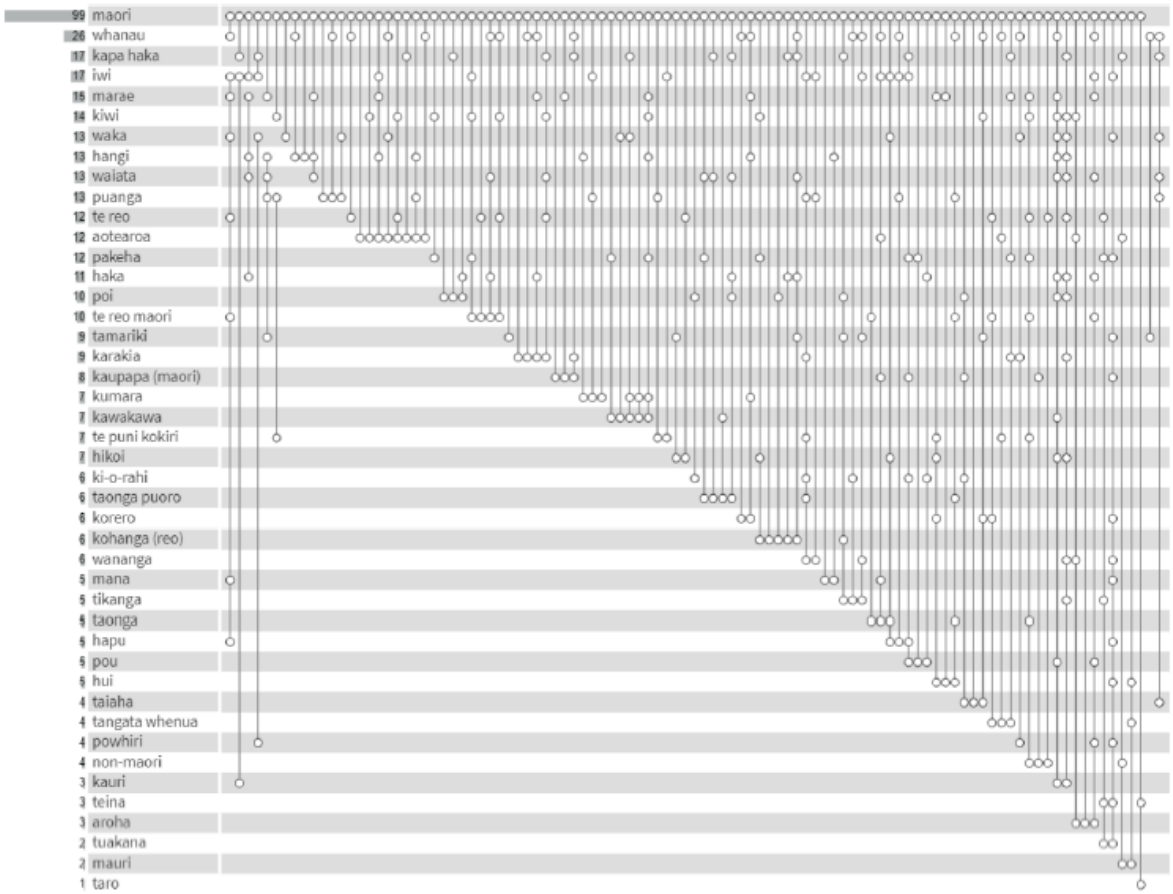

Figure 21. All 125 sets in the corpus, including the outlier *Māori*