Visualising Categorical Data: Linguistic Case Studies from te Reo Māori and New Zealand English

A thesis

submitted in fulfilment

of the requirements for the Degree

of

Doctor of Philosophy

at

The University of Waikato

by

David Gareth Trye

THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2024

Abstract

Categorical variables are prevalent in real-world datasets across numerous domains, yet few visualisation techniques accommodate them effectively. This is especially true of datasets comprising three or more categorical variables, termed multivariate categorical data. Visualising such data is challenging due to the lack of inherent ordering of nominal categories, the so-called 'curse of dimensionality', and the potential variability in the number of categories per variable. Corpus linguistics, which involves the study of large digital collections of naturally occurring language, serves as the primary application domain in this thesis. This domain was chosen because it is rich in multivariate categorical data and, at the same time, is often visualised using only basic techniques.

This thesis contributes to the area of categorical data visualisation in several ways. First, we propose a taxonomy of techniques for visualising categorical data, highlight limitations of existing solutions and identify relevant analysis tasks. Building on this foundation, the thesis introduces novel techniques and enhancements for visualising datasets involving multiple categorical variables. We focus on adapting the layout and interactive capabilities of an existing technique that uses a matrix of heatmaps to represent pairwise category intersections. These modifications show that directly visualising statistical test results for categorical data can be beneficial for exploring bivariate patterns and associations. Furthermore, we contribute the design, implementation and evaluation of a novel technique called MultiCat, which is not restricted to pairwise intersections but rather facilitates analysis of relationships among multiple variables simultaneously. Both these techniques are interactive and offer greater scalability than existing alternatives, thereby affording new possibilities for analysing multivariate categorical data. However, since categorical variables can occur within more complex data structures, we also consider their presence in networks and hypergraphs, which require specialised methods.

To demonstrate the application of these techniques, we draw on two linguistic case studies that focus on languages of special significance in Aotearoa New Zealand. Addressing the low-resource status of Māori, the country's Indigenous language, we first contribute two related Twitter datasets—a mono-

lingual Māori corpus and a mixed–language Māori–English corpus—together with an architecture for differentiating Māori and English words. Our initial case study uses the monolingual Māori corpus and proposed visualisation techniques to investigate grammatical possession in Māori, offering fresh insights into the linguistic practices of contemporary speakers. The second case study uses networks and hypergraphs with categorical attributes to explore Māori loanword co-occurrence in New Zealand English newspaper articles. We find that loanwords tend not to occur in isolation and that New Zealanders are still importing new (unlisted) borrowings from Māori.

Ultimately, the techniques developed in this thesis have broad applications both within and beyond the corpus linguistics community. By enabling more effective visualisation and analysis of multivariate categorical data, this research has the potential to facilitate deeper insights into domains as diverse as education, healthcare, business and science.

Acknowledgements

Over the past four years, my standard response to the question 'What does PhD stand for?' has been 'Please help David!', said with increasing degrees of desperation. Fortunately, I have received a great deal of help from a great many people, to whom I am incredibly grateful.

Firstly, I am privileged to have co-authored papers with nine people as part of this research: thank you all! In particular, I would like to thank my wonderful supervisors, Mark Apperley, David Bainbridge and Andreea Calude, whose influence extends far beyond these pages. Mark, I deeply appreciate your wisdom, patience and unwavering support. Thank you for giving me the freedom to develop ideas at my own pace, and for subtly steering me in the right direction whenever I veered off course. I shall never forget, thanks to you, that an informative visualisation—and indeed a thesis—should always tell a story. David, thank you for your diligent reading of drafts and your kind reassurance. I have greatly enjoyed our lively discussions, your humorous anecdotes and analogies. Your 'Bainbridge Brackets' rule has undoubtedly reduced the number of parentheses in this thesis. Thanks also for helping me to overcome my fear of LaTeX, one 'Recompile' at a time. Andreea, thank you for taking me under your wing while I was an undergraduate student and for always being in my corner. You inspired me to embark on this journey, and I am grateful to have had the opportunity to integrate linguistics into my research. What is language if not a never-ending quest for knowledge?

Thank you to the many staff from the School of Computing and Mathematical Sciences whom I had the pleasure of working with in my capacity as Doctoral Assistant, especially Tim Elphick, Nilesh Kanji, Sunitha Prabhu and Phil Treweek. My heartfelt thanks go to Te Taka Keegan, Jemma König, Cameron Grout, Vithya Yogarajan, Ray Harlow, Felipe Bravo-Marquez, Alvin Yeo and Leigh Sanderson for their mentorship and generosity over the years. Ngā mihi nui ki a koutou katoa. Thanks also to my colleagues on the Student Discipline Committee, who often joked that they hoped this thesis would never see the light of day, lest they have to replace me. I like to think that was because I did a good job and not because they thought no one else would be interested in filling the position.

I gratefully acknowledge the University of Waikato for funding my research through a Doctoral Scholarship, which made this endeavour possible. Thanks also to my supervisors for supporting my trip to Finland, at a time when attending a conference on the other side of the world seemed like only a remote possibility (if you'll excuse the pun).

Next, a huge thank you to my friends for always knowing what to say to cheer me up, providing much needed laughter and distractions from my work, and putting up with my many idiosyncrasies. You know who you are!

Last but not least, I would like to thank my family for their unconditional support throughout this journey. To my parents, Diana and Keith, I could not have done this without your love and encouragement, especially through some of the toughest stretches of this work. Thank you for fuelling my chocolate addiction, enduring my procrastination rituals and telling everyone who would possibly listen that my PhD has 'something to do with computers and language'. To my sisters, Suzanne and Marianne, thank you for checking in on me, and for your kind and encouraging words. Most importantly, thank you all for believing that I was capable of completing this thesis, even when I occasionally stopped believing in myself. Finally, an honourable mention to Katy Cat, who frequently reminded me of the importance of backing up my work to safeguard against a keyboard cat-astrophe.

Contents

List of Figures

List of Tables

Part I

Thesis Preliminaries

Chapter 1

Introduction

As datasets grow in size and complexity, more specialised visualisation techniques are needed to facilitate their effective exploration. However, techniques designed for categorical data—spanning both nominal and ordinal variables—have received little attention in the visualisation literature compared to those for continuous data, especially in contexts involving multiple variables. Such datasets arise in diverse domains: for instance, census data typically include variables such as gender, education level, religion and marital status; medical records might include disease types, treatment protocols and patient outcomes; retail databases frequently categorise products by type, payment method and customer demographics. Analysing all categorical variables simultaneously in these datasets can enhance understanding of complex relationships and support informed decision-making. This thesis contributes to this gap in existing visualisation techniques by proposing novel solutions that can effectively accommodate larger numbers of categorical variables. The rest of this chapter will define the main research questions, provide an overview of the thesis structure, summarise key contributions, and outline the publications and manuscripts that are included as separate chapters in the thesis.

1.1 Research Questions

This thesis seeks to answer the following fundamental questions:

1. What generalisable information visualisation techniques can be developed or adapted to enable the effective analysis of datasets involving multiple categorical variables?

2. How can applying these techniques to a particular domain increase understanding of that domain?

We will apply the techniques resulting from addressing the first research question to the domain of corpus linguistics in order to address the second. While the presented examples centre on datasets related to Māori and New Zealand English (NZE), both of which are significant to Aotearoa[1] New Zealand, the techniques are generalisable. We bring out this aspect of the work in discussion sections throughout the thesis.

1.2 Thesis Structure

The content of this thesis is divided into four parts and ten chapters, as shown in Figure 1.1. These chapters span the overlapping fields of *Information Visualisation*, *Natural Language Processing (NLP)* and *Corpus Linguistics*, reflecting the interdisciplinary nature of the research carried out.

Part I includes this introduction (Chapter 1), together with relevant background details on information visualisation, corpus linguistics and the main languages of interest (Chapter 2).

Part II focuses on visualisation techniques that support purely categorical data, rather than mixed or continuous data types. We begin with a review and taxonomy of established techniques for visualising categorical data (Chapter 3), before proposing extensions to a recent technique for visualising pairwise category intersections (Chapter 4). We then introduce and evaluate a novel technique for visualising higher-order categorical relationships (Chapter 5).

Part III is concerned with linguistic case studies from te reo Māori (the Māori language) and New Zealand English. Motivated by the lack of existing datasets for Māori, we created two social media corpora that provide rich opportunities for visualising categorical data (Chapters 6–7). This is followed by a case study on grammatical possession in Māori (Chapter 8), which applies the techniques from Part II to the dataset introduced in Chapter 6. We then examine how Māori loanwords co-occur in New Zealand newspaper articles, expanding our problem space from purely categorical data to relational data (networks) with categorical attributes (Chapter 9).

Finally, **Part IV** (Chapter 10) provides a summary of the main findings and contributions of this thesis, and outlines avenues for future work.

[1] *Aotearoa* is the Māori name for New Zealand, and will be used throughout this thesis.

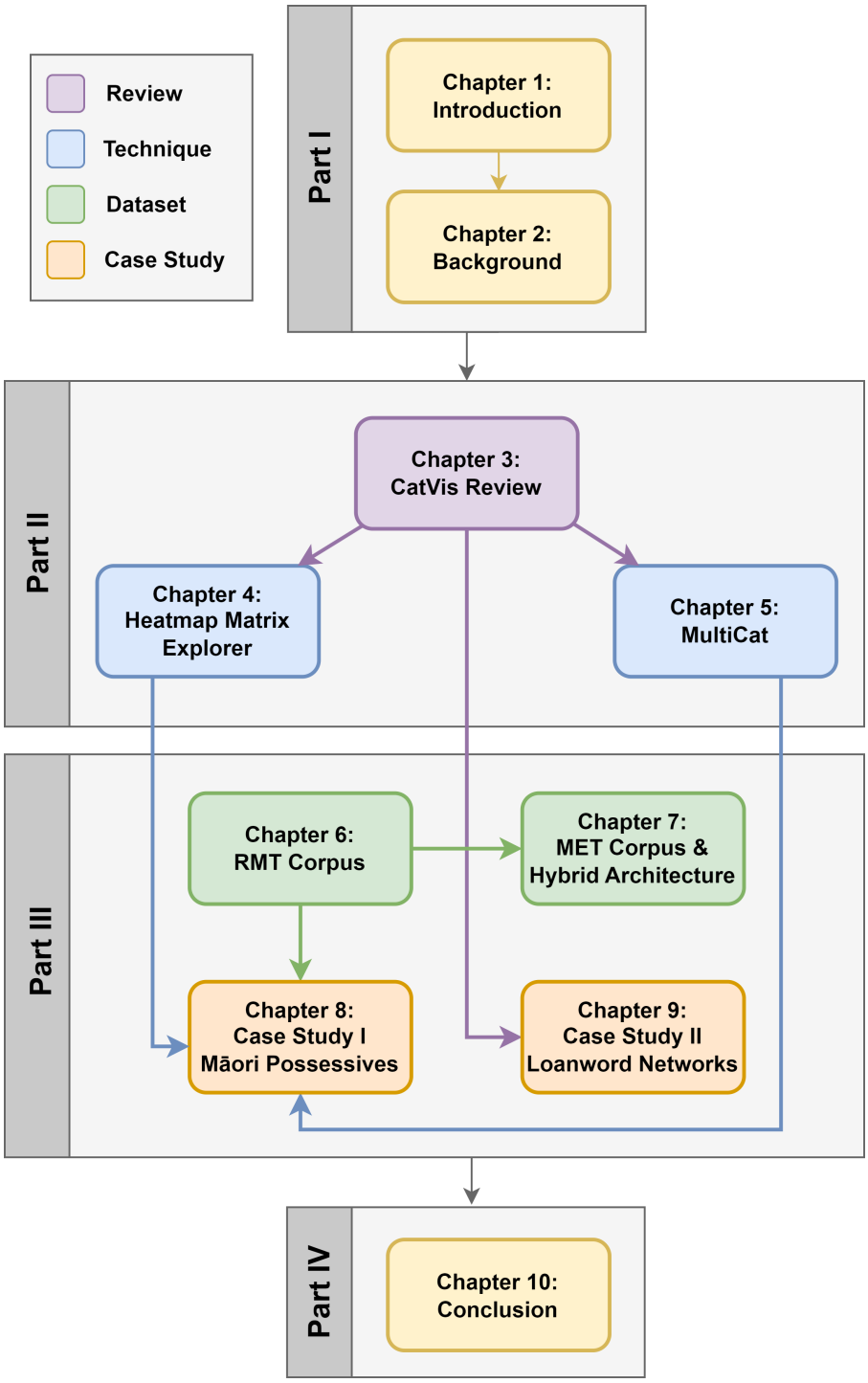


Figure 1.1: Overview of the structure of this thesis and the main links between chapters.

1.3 Contributions

The contributions made by this thesis can be described as follows:

1. **A review of categorical data visualisation**, including a taxonomy of techniques and an overview of relevant analysis tasks (Chapter 3).

2. **New and adapted visualisation techniques** for exploring multidimensional categorical data (Chapters 4–5) and relational data with categorical attributes (Chapter 9). These techniques advance the state of the art in categorical data analysis and can be generalised to domains such as business, science, education and communication.

3. **Language resources** for te reo Māori and the mixing of Māori and English, including two related Twitter corpora and an architecture for labelling additional bilingual Māori/English datasets (Chapters 6–7).

4. **Linguistic findings** about how te reo Māori and New Zealand English are used in contemporary New Zealand society (Chapters 8–9), obtained by applying the aforementioned visualisation techniques.

1.4 List of Publications and Manuscripts

Chapters 4–9 of this thesis comprise published papers or manuscripts prepared for publication, as follows:

- Chapter 4 (published): **Trye, D.**, Apperley, M, & Bainbridge, D. (2023). Extending the Heatmap Matrix: Pairwise analysis of multivariate categorical data. In *2023 27th International Conference Information Visualisation (IV).* (pp. 29-36). Tampere, Finland: IEEE. `https://doi.org/10.1109/IV60283.2023.00016`

- Chapter 5 (to be submitted): **Trye, D.**, Apperley, M., & Bainbridge, D. (2024). *MultiCat: A visualisation technique for multidimensional categorical data* [Unpublished manuscript]. School of Computing and Mathematical Sciences, University of Waikato.

- Chapter 6 (published): **Trye, D.**, Keegan, T. T., Mato, P., & Apperley, M. (2022). Harnessing Indigenous Tweets: The Reo Māori Twitter Corpus. *Language resources and evaluation*, *56*(4), 1229-1268. `https://doi.org/10.1007/s10579-022-09580-w`

- Chapter 7 (published): **Trye, D.**, Yogarajan, V., König, J., Keegan, T. T., Bainbridge, D., & Apperley, M. (2022, November). A hybrid architecture for labelling bilingual Māori-English tweets. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (pp. 119-130). https://aclanthology.org/2022.findings-aacl.11

- Chapter 8 (published): **Trye, D.**, Calude, A. S., Harlow, R., & Keegan, T. T. (2024). Analysing A/O possession in Māori-language tweets. *Languages*, *9*(8), 271. https://doi.org/10.3390/languages9080271

- Chapter 9 (published): **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

The five published chapters appear in peer-reviewed journals (Chapters 6, 8 & 9) and conference proceedings (Chapters 4 & 7), with the writing style reflecting the intended audience. All included publications and manuscripts have been reproduced with minor changes to the formatting for consistency throughout the thesis. In particular, adjustments have been made to the layout and pagination of each chapter, including the renumbering of all sections, figures and tables, as well as the re-lettering of appendices and supplementary material. Additionally, all references have been standardised to follow a style in line with APA. In some chapters, hyperlinks that originally appeared as footnotes have been converted to in-text citations. Any further deviations from the source material are noted in context. Because each of these chapters has been written as a stand-alone publication, there is inevitably some repetition between chapters.

In order to reinforce the links between chapters and the broader thesis, Chapters 2–9 each begin with an introduction and conclude with a postscript. These postscripts are immediately followed by references for the corresponding chapter. For ease of readability, appendices are not included within the chapters themselves; instead, they are grouped at the end of the thesis in the order in which they are cited.

This research is the outcome of fruitful collaboration with many different people, including my PhD supervisors, mentors and colleagues at the University of Waikato (see Appendix A for details). In recognition of their collective contribution, I will use the first-person plural throughout the thesis.

Chapter 2

Background

The research presented in this thesis is interdisciplinary, combining ideas and methodologies from the fields of Information Visualisation, Natural Language Processing and Corpus Linguistics. We therefore provide some background information about these disciplines, together with important context for understanding our case studies in Part III.

2.1 Information Visualisation Fundamentals

Visualisation is commonly defined as "the use of computer-supported, interactive, visual representations of data to amplify cognition" (Card et al., 1999, p. 6). These representations facilitate the exploration, confirmation or presentation of data, leading to insights that might otherwise go unnoticed. Importantly, visualisations leverage the advanced perceptual capabilities of the human visual system, whose bandwidth surpasses all other senses combined (Ware, 2019, p. 2). *Interactive* visualisations are particularly valuable as they enable users to navigate through a dataset according to their specific needs, thereby harnessing the complementary strengths of humans and computers.

A distinction is often made between Information Visualisation (*InfoVis*), which deals with *abstract* data, and Scientific Visualisation (*SciVis*), which focuses on *physical* data. This thesis is concerned with the former.

2.1.1 The Visualisation Pipeline

While computer-based visualisations typically have unique characteristics, they can be systematically analysed using the Visualisation Pipeline (Card et al., 1999, p. 17), shown in Figure 2.1. This pipeline describes the mapping of data to visual form to support human interaction. The first stage involves preprocessing and transforming the raw data, by extracting data from source

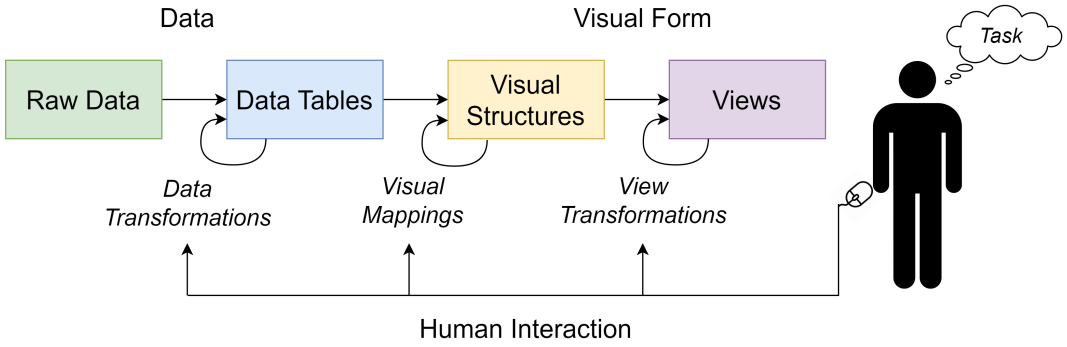


Figure 2.1: The Visualisation Pipeline, redrawn and adapted from Card et al. (1999).

files, removing extraneous information, converting the data into a particular format for use in a visualisation tool, interpolating missing values, and so on. Crucially, the way in which the data is preprocessed affects what exploration is possible later on. The second stage is the visual mapping process, during which a visual encoding is specified that determines how the data will be represented (discussed next). Finally, view transformations, such as zooming and filtering, affect which parts of the data are rendered on the screen. The feedback loops stemming from the user indicate that each of these steps can be continually refined.

2.1.2 Representation

The two basic building blocks of visualisation are *representation* and *interaction*. Representation is concerned with transforming data into visual form, which is achieved by mapping data attributes to *visual channels*, also called *visual variables*. There is a finite set of visual channels available, including position, size, texture, colour, shape and, in computer-based visualisation, motion or animation (Bertin, 1983; Carpendale, 2003). An effective example of a visualisation that leverages multiple visual channels and comprises mixed data types is given in Figure 2.2, which plots life expectancy against average daily income. The size of the circle for a country is proportional to its population, and the colour used denotes the continent to which it belongs. The same visual channel should never encode more than one variable, but the same variable may be redundantly encoded by multiple visual channels. In Figure 2.2, for example, colour should not be used to encode both the continent of the country and its population, but the continent could be encoded using both colour and texture, given that texture has not already been used elsewhere.

The appropriate encoding mechanism for a given context depends on the

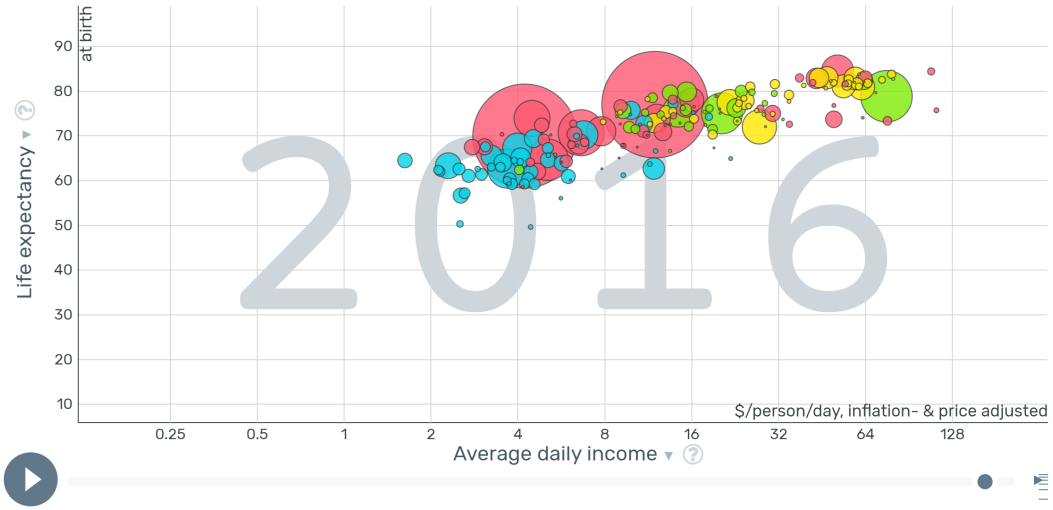


Figure 2.2: A motion chart displaying countries' *Average daily income* (x-axis), *Life expectancy* (y-axis), *Population* (size) and *Continent* (colour), animated across *time* (screenshot taken from `https://www.gapminder.org`).

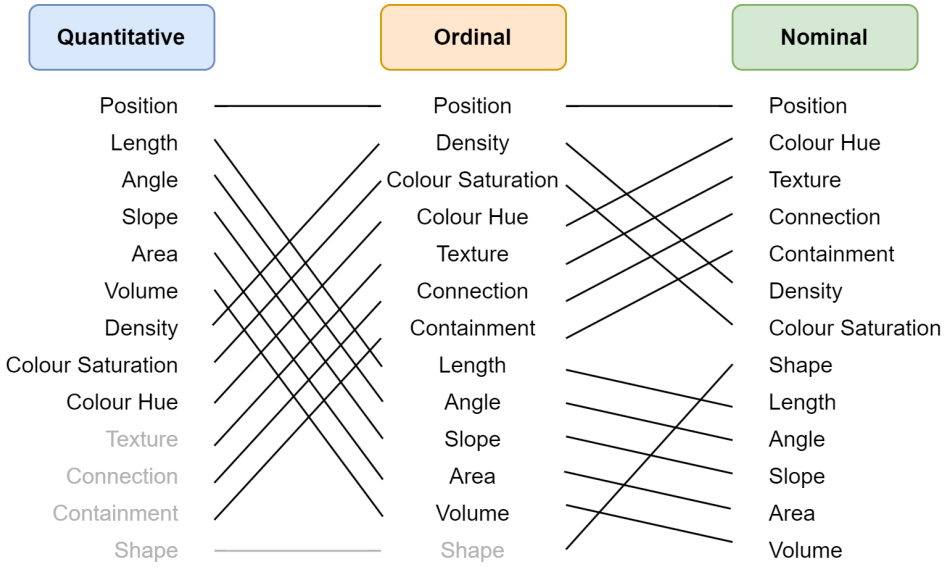


Figure 2.3: Mackinlay's ranking of visual channels (which he refers to as 'tasks') for different data types, ordered from most effective at the top to least effective at the bottom. Visual channels shown in grey are not relevant to the corresponding data type (redrawn and adapted from Mackinlay, 1986).

characteristics of the data: some visual channels are better suited to quantitative, ordinal or nominal variables. This thesis focuses on the latter two data types, which are collectively known as categorical data. Building on work by Cleveland and McGill (1984), Mackinlay (1986) ranked the accuracy of the different visual channels for each of these data types, as shown in Figure 2.3. For example, it is evident from this figure that position is considered the most accurate encoding for all three data types, colour is more suited to nominal data than other data types, and length is always more accurate than area. The choices of visual encoding significantly affect the perception of the data, as well as the ability to make valid comparisons. As such, when creating any visual data representation, the most important attributes in a dataset should always be ascribed to the most salient visual channels; this is known as the *effectiveness principle* (Munzner, 2014, p. 101).

2.1.3 Perceptual Guidelines

It is important to design visualisations by considering the human capacity for visual perception and cognition (Franconeri et al., 2021). Pre-attentive processing is a special form of perception that can be used, among other things, to quickly identify *distinct elements* in complex arrangements (Healey et al., 1996). Similarly, Gestalt theory provides robust guidelines for arranging *groups* of elements based on how humans perceive visual patterns (Koffka, 1935; Ware, 2019). For example, according to the Gestalt Law of Similarity, items with similar visual attributes tend to be perceived as a group, which is exemplified in Figure 2.4. Here, colour is used to 'group' bars belonging to the same category (corpus), enabling quick and effective comparisons. If visual channels are the main way of describing data, the Gestalt Laws can be seen as a secondary layer that provide additional context, by grouping or partitioning elements in meaningful ways.

2.1.4 Interaction

Visualisations can be greatly enhanced through interaction. Metaphorically speaking, interaction allows users to change the lens on the data. By manipulating a visualisation, users can observe cause-and-effect relationships and gain confidence in formulating and answering their own questions, thereby facilitating information acquisition (Tominski, 2022). Interaction is helpful for exploring large and complex datasets because (i) there is often too much information to display at once, and (ii) there are many different ways it can be

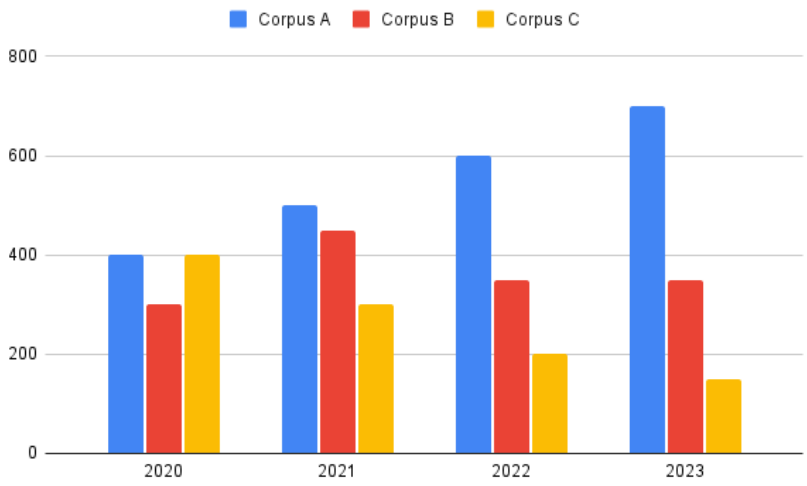


Figure 2.4: Grouped bar chart illustrating the Gestalt Law of Similarity through the use of colour.

presented, only some of which will support a user's specific needs. Interaction is particularly valuable when it allows the user to drill down into a dataset in a flexible and open-ended manner (Heer and Shneiderman, 2012).

There are a variety of high-level tasks that reflect users' motivations and interests when they are engaged with a visualisation. Shneiderman's (1996, p. 337) classic 'Visual Information Seeking Mantra' constitutes a useful starting point for thinking about interaction and the sorts of tasks it can support: "Overview first, zoom and filter, then details on demand". In the same paper, Shneiderman mentions three further tasks: 'relate' (view relationships among items), 'history' (keep a history of actions to support undo, replay, and progressive refinement) and 'extract' (allow extraction of sub-collections and of the query parameters). Yi et al. (2007, p. 1226) define a similar, updated list of high-level tasks that are also motivated by users' intentions:[1]

- Select: mark something as interesting.
- Explore: show me something else.
- Reconfigure: show me a different arrangement.
- Encode: show me a different representation.
- Abstract/Elaborate: show me more/less detail.
- Filter: show me something conditionally.
- Connect: show me related items.

[1]The bulleted task descriptions in this section have been collated from the source material, using the original wording.

Interaction can also be conceptualised in terms of low-level tasks that are centred around the *data* rather than the user or their intentions. Amar et al. (2005, p. 113) provide a list of such tasks:

- Retrieve Value: Given a set of specific cases, find attributes of those cases.
- Filter: Given some concrete conditions on attributes values, find data cases satisfying those conditions.
- Compute Derived Value: Given a set of data cases, compute an aggregate numeric representation of those data cases.
- Find Extremum: Find data cases possessing an extreme value of an attribute over its range within the dataset.
- Sort: Given a set of data cases, rank them according to some ordinal metric.
- Determine Range: Given a set of data cases and an attribute of interest, find the span of values within the set.
- Characterise Distribution: Given a set of data cases and a quantitative attribute of interest, characterise the distribution of that attribute's values over the set.
- Find Anomalies: Identify any anomalies within a given set of data cases with respect to a given relationship or expectation, e.g. statistical outliers.
- Cluster: Given a set of data cases, find clusters of similar attribute values.
- Correlate: Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.

Conceptually, these low-level tasks answer the 'how' of interaction, whereas the high-level tasks answer the 'why' (Brehmer and Munzner, 2013). Common interaction techniques that support these tasks include: focus+context techniques (Cockburn et al., 2009), brushing and linking (Becker and Cleveland, 1987); dynamic queries (Shneiderman, 1994), semantic zoom (Bederson and Hollan, 1994); and scrolling (e.g., Seyser and Zeiller, 2018). Figure 2.5 is an example of brushing and linking, where data items selected in one view are highlighted in the other.

2.1.5 Multidimensional Visualisation

The representation of multidimensional data poses a significant challenge in the field of information visualisation. A dataset is considered multidimensional if it comprises three or more variables, which can be quantitative, ordinal or

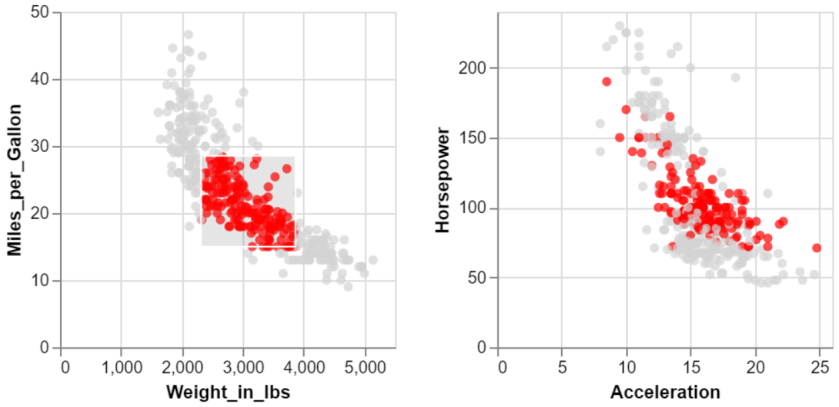


Figure 2.5: Brushing and linking across two scatter plots (screenshot taken from `https://vda-lab.github.io/visualisation-tutorial/vegalite-brushing-and-linking.html`).

nominal. More precisely, each data item in a multidimensional dataset can be defined as $d_i = (d_{i1}, d_{i2}, \ldots, d_{iN}), i \in \{1, 2, \ldots, M\}$, where i is the item's index, N is the number of variables (with $N \geq 3$), and M is the total number of data items. Multidimensional datasets are commonly structured as tables, where rows correspond to data items and columns correspond to variables. We will use the terms 'multidimensional' and 'multivariate' interchangeably throughout this thesis.

A wealth of multidimensional visualisation techniques have been proposed over the past few decades, some of which are more intuitive than others (Liu et al., 2016). These techniques aim to facilitate exploration of a large number of variables at the same time, allowing users to find meaning in the diversity, even if they do not start out with a clear idea of what to look for (Ware, 2019). Regardless of the number of variables they represent, visualisations are typically rendered in two- or three-dimensional space, due to unavoidable cognitive limitations.

Visualisation techniques for representing multidimensional data can be broadly classified into the following five categories (Keim, 2000, p. 60):

- Geometric techniques, which use position and/or size as the main visual channels to encode the data, such as scatter plots, scatter plot matrices (SPLOMs), parallel coordinates (Inselberg, 1985) and *Table Lens* (Rao and Card, 1994).
- Icon-based techniques, which typically map individual data items to distinct icons or glyphs, such as *Chernoff Faces* (Chernoff, 1973) and star plots (Coekin, 1969).
- Pixel-oriented techniques, where data values are encoded using coloured

pixels, such as *Circle Segments* (Ankerst et al., 1996) and the *Recursive Pattern* (Keim et al., 1995).

- Hierarchical techniques, where the data are recursively divided to show the values of multiple variables, such as mosaic plots (Hartigan and Kleiner, 1981) and dimensional stacking (LeBlanc et al., 1990).
- Graph-based techniques, which model relationships between data items and variables using nodes and edges, such as *PivotSlice* (Zhao et al., 2013) and *Ploceus* (Liu et al., 2011).

These categories are not mutually exclusive, and additional methods are available for displaying multiple variables simultaneously. For example, Chernoff Faces can be integrated into a two-dimensional scatterplot, resulting in a hybrid geometric/icon-based representation. Furthermore, Multiple (Coordinated) Views (Roberts, 2007) facilitate the effective juxtaposition of visualisations encoding separate dimensions, while potentially also leveraging the strengths of different types of representations. Additionally, projection methods—such as Principal Component Analysis (PCA), Correspondence Analysis (CA) and Multidimensional Scaling (MDS)—are frequently employed to reduce the number of variables before the data are mapped to visual form, often using geometric techniques.

There are, however, a number of unsolved problems regarding the visualisation of multidimensional data, including complex categorical data, as discussed in Chapter 3. The *curse of dimensionality* (Bellman, 1961) means that data become increasingly sparse as the number of variables increases, making many types of data analysis more difficult. At the same time, many existing visualisation techniques are prone to clutter and/or visual occlusion, limiting their scalability.

Datasets containing many variables are often too large or complex to be visualised in a straightforward manner, unless some of the detail is abstracted away, thereby sacrificing granularity for greater readability. In order to minimise loss of critical information and provide a helpful solution, it is necessary to have an in-depth understanding of not just the *data*, but also the *users* and the *tasks* they wish to perform. For instance, consider a social scientist analysing demographic data from thousands of survey respondents, including categorical variables such as gender, education level, political affiliation and country of residence. The researcher needs to identify over-represented groups of respondents based on these categories. To accomplish this task quickly and efficiently using visualisation, data items with identical values would need to be aggregated rather than displayed individually.

2.2 Corpus Linguistics

Having introduced key ideas from information visualisation, this section explains why corpus linguistics is a useful target domain for exploring multidimensional categorical data. We begin with an overview of corpus linguistics and highlight the current state of visualisation techniques within the field.

Over the past forty years, theoretical linguistics has increasingly adopted empirical, data-driven methods alongside traditional, introspective ones. This shift has been accompanied by the rise of corpus linguistics, which focuses on analysing large corpora of naturally occurring text (McEnery and Hardie, 2011). These corpora are designed to be representative of typical sociolinguistic variables such as time, genre, or the social status of each speaker. Leveraging data-driven analysis, scholars can gain insights into a wide range of linguistic phenomena, including lexical, morphological, syntactic and semantic features that would otherwise be difficult to uncover (Biber et al., 1998).

A variety of text analysis methods have flourished in corpus linguistics. For instance, frequency profiling involves counting how often words, phrases or other linguistic elements appear in a corpus, while concordances provide a means of examining co-occurrence patterns, revealing grammatical and usage norms. These methods have been widely implemented in corpus analysis software: popular examples include *AntConc* (Anthony, 2023), *English-Corpora.org* (Davies, 2020), *#LancsBox X* (Brezina and Platt, 2024), *Sketch Engine* (Kilgarriff et al., 2014) and *WordSmith Tools* (Scott, 2024).

2.2.1 Visualising Language Data

Linguistic visualisation concerns "the presentation of linguistic data through visual representations designed to amplify cognition or communicate linguistic information" (Collins, 2010, p. 44). In the field of corpus linguistics, where the complexity and volume of data are continually increasing, visualisations are not just beneficial but—in many cases—necessary for exploring intricate patterns, explaining phenomena and facilitating statistical analysis (Siirtola et al., 2014).

There are two possible approaches for developing tools that support linguistic visualisations. The first is to create specialised applications that are linguistically motivated (see Butt et al., 2020 and references therein). Prominent examples include *Diachronlex diagrams* (Theron and Fontanillo, 2015), the *Text Variation Explorer* (Siirtola et al., 2014), *Concordance Mosaics* (Sheehan et al., 2022) and *AppAnn* (Almutairi, 2013), as well as several tools de-

veloped by Chris Culy, which are available online.[2] Recent advances have also seen the incorporation of machine learning in linguistic tools to enhance the analyst's workflow (Schneider et al., 2017), along with visualisation tools that integrate with existing corpus analysis software (Isaacs et al., 2024).

The second approach—and the one adopted in this thesis—is to design more general (domain-agnostic) visualisation tools or techniques, to which language data can be directly applied (e.g., Siirtola et al. 2011; Hilpert, 2011). Tools like *Mondrian* (Theus, 2002) and *Tableau* (Stolte et al., 2008), for instance, can be useful for experimenting with techniques that have not been implemented in specialised linguistic tools. Regardless of which approach is used, no single visualisation technique can fully accommodate every linguist's needs; it is therefore helpful to have a diverse inventory from which to choose.

Information visualisation poses an ongoing methodological challenge for corpus linguists, who have generally been slow to adopt new visualisation techniques (Isaacs et al., 2024, p. 1; Anthony, 2018, p. 198). In fact, the structure and content of corpora are typically described using words and tables rather than visualisations (Sönning and Schützler, 2023). Given the advantages that visualisation has to offer, Sönning and Schützler (ibid, p. 12) describe this as "somewhat surprising, if not unsatisfactory". Their analysis of 1,238 corpus linguistics articles from 2015 to 2020 revealed that almost a third of these papers did not contain any visualisations at all (though, of course, this does not necessarily mean that the authors did not create or interact with any visualisations when undertaking the research). Among the remaining papers, authors displayed a strong preference for a small number of basic chart types, including bar charts, line plots, scatter plots and dot plots (see also Anthony, 2018, p. 202, Allen, 2017, pp. 464–465; Rayson et al., 2016, p. 28). As Anthony notes (2018, p. 198), these and other basic charts are not always suitable for capturing the complexity of the underlying data, which calls for more sophisticated visualisation techniques. The lack of variation in the visualisations chosen by corpus linguists can be at least partly attributed to current limitations of corpus analysis tools, which do not tend to support more advanced techniques (ibid, p. 207). Overall, these observations suggest that visualisations are under-utilised by the corpus linguistics community. Hence, it is worthwhile increasing awareness of relevant visualisation techniques and ensuring they are easily accessible to corpus linguists.

There are several reasons why corpus linguists can benefit from more advanced visualisation techniques. Firstly, exploratory visualisation fits very well

[2] `http://linguistics.chrisculy.net/lx/software/`

within the corpus linguistics methodology, as discussed below. Secondly, the size and availability of digital corpora is growing, providing rich resources for analysis, but simultaneously making it harder to extract meaningful patterns from the data. Thirdly, there is a growing interest in computer-mediated language, particularly language produced on social media and the internet, which introduces additional complexities (Calude, 2023). Robust visualisation tools are needed to enable linguists to extract relevant features quickly and efficiently to support their analyses (Rayson et al., 2016; Anthony, 2018). Importantly, however, these visualisations should "aid rather than replace linguists' own expertise in making sense of real world language" (Allen, 2017, p. 478).

Interactive visualisations are extremely valuable, yet surprisingly rare within corpus linguistics. Rayson et al. (2016, p. 34) note that many existing tools are static in nature, often resulting in visual clutter as more information is presented than desirable. They argue that more dynamic, interactive and iterative visualisation tools are needed to support the data-driven corpus methodology. Similarly, Siirtola et al. (2014, p. 11) advocate greater use of exploratory techniques in corpus linguistics, rather than simple text concordances and spreadsheet applications, especially when users are in the initial stages of an analysis. The shortcomings of existing tools highlight the value of developing novel interactive solutions that foster more playful exploration of linguistic data and more efficient data analysis. Linguists are more likely to invest time into learning visualisation tools if they know they are easy to use and can reward them with additional opportunities for insight.

2.2.2 Relationship to Text Visualisation

Text visualisation is an important subfield of information visualisation, which has received considerable attention in recent years (Kucher and Kerren, 2015; Jänicke et al., 2015; Brath, 2020). Given that corpora are made up of text, it would be natural to assume that corpus visualisations necessarily involve the representation of textual data. However, while these two areas overlap, not all visualisations in corpus linguistics directly represent the text within a corpus. For example, linguistic analyses often abstract textual data into categories (e.g., syntactic classes, phonemic categories), thus distancing the representation from the original text (see the next section). Moreover, text visualisation extends beyond linguistics to encompass broader applications in the social sciences and digital humanities. Consequently, many techniques developed for visualising text are not tailored to the needs of corpus linguists (Siirtola et al., 2016; Culy and Lyding, 2010).

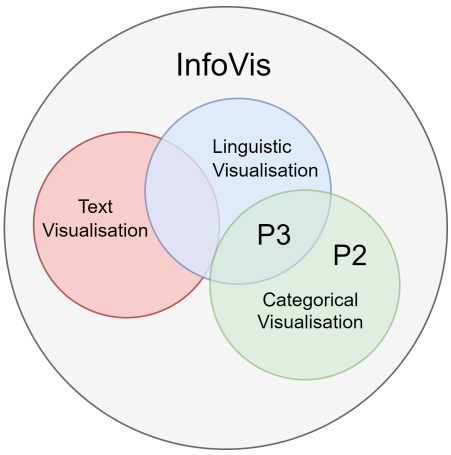


Figure 2.6: Venn diagram showing the relationship between text visualisation, linguistic visualisation and categorical visualisation ('P2' and 'P3' designate Part II and Part III of this thesis, respectively).

The relationship between text visualisation, linguistic visualisation and categorical visualisation, as far as this thesis is concerned, is illustrated in Figure 2.6. We focus on categorical visualisation techniques, represented by the green circle, rather than on text visualisation. Part II of the thesis looks at the general area of categorical visualisation, with Part III looking specifically at the intersection of this area with linguistic visualisation.

2.2.3 Visualising Multidimensional Categorical Data

As this thesis explores visualisation techniques for categorical data, and specifically *multidimensional* categorical data, our target domain needs to provide opportunities for visualising numerous categorical variables. Corpus linguistics meets these criteria because (i) language data is inherently multidimensional, and (ii) categorical variables are prevalent in corpus linguistic studies. Regarding this first point, Almutairi (2013, p. 698) observes that "The multidimensional nature of textual data is a challenging problem and most linguists are interested in understanding more than three or four dimensions in their textual data simultaneously." This renders many existing visualisation techniques unreadable, especially those designed to display the frequency of all category intersections simultaneously.

Regarding the second point, categorical variables are the most frequent data type in corpus linguistics (Stefanowitsch, 2020, p. 177). The sheer dominance of bar charts in corpus linguistics publications, and the increasing popularity of mosaic plots (Sönning and Schützler, 2023), also provide evidence of this. Common examples of categorical variables include *sentiment* (positive,

neutral, negative), *intensity* (low, medium, high), *position* (beginning, middle, end), *animacy* (human, animate, inanimate, abstract) and *construction type* (e.g., ditransitive, prepositional dative). Moreover, categorical 'background' variables of the speakers or writers of each text are also frequently considered in analyses, such as their *gender*, *ethnicity* and *socio-economic status* (Siirtola et al., 2011).

The visualisation techniques considered in this research will be of particular relevance to corpus linguists who commonly analyse specific features or constructions by tagging them according to large numbers of categorical variables, whether this tagging is done manually, automatically or semi-automatically.[3] Two recent examples of datasets structured this way are Burnette and Calude (2022) and Calude and Delahunty (2020), which each contain 10 categorical variables. The techniques proposed in this thesis are based on the assumption that the main variables of interest are categorical and that analysts will want to view them all simultaneously—an aspect not well supported by current approaches, as will be discussed in Part II.

2.3 Te Reo Māori and New Zealand English Context

We now provide some context about the two languages of interest in this thesis, namely te reo Māori (the Māori language) and New Zealand English (NZE), both of which will be explored in Part III.[4] These languages were selected because of their relevance to the local research setting in Aotearoa New Zealand: Māori is a *de jure* official language of New Zealand (alongside New Zealand Sign Language), while English is *de facto* official.[5] Although the vast majority of New Zealanders are English monolinguals, Māori plays an increasingly prominent role in the country's life, and a more inclusive attitude towards the language appears to be emerging (Albury, 2015, 2016; Te Puni Kōkiri, 2010; Hashimoto, 2019). Indeed, the Māori language is special to Aotearoa as it provides a unique context to experience the country's Indigenous culture and history, and to understand the values and worldview of the Māori people.

[3] Automated procedures can sometimes generate additional categorical variables at little extra cost.

[4] In particular, we will focus on Twitter data containing varying degrees of Māori (Chapters 6–8), and on NZE newspaper articles featuring Māori loanwords (Chapter 9).

[5] While we focus only on these two languages, we note that more than 160 languages are used in New Zealand.

2.3.1 Te Reo Māori

Māori is the Indigenous language of Aotearoa New Zealand, and is a member of the East Polynesian subgroup of the Austronesian language family.[6] The language is essential to the creation and expression of Māori identity (Marras Tate and Rapatahana, 2022). It has a small phoneme inventory comprising ten consonants /p, t, k, m, n, ŋ, w, f, ɾ, h/ and five vowels /i, e, a, o, u/ with long vowel pronunciation indicated using a macron. In terms of syntax, Māori has basic VSX (verb-subject-object) order in simple sentences; however, variations exist for fronting topical or focal phrases (Harlow, 2007). Māori is not a single homogeneous language but rather comprises a small number of mutually intelligible regional dialects, called *mita* (Biggs, 1968). Much of the variation across *mita* is lexical, but there are also some phonological differences (Harlow, 2007). Although Māori has a rich history of language description (see Whaanga and Greensill, 2014), including several grammars (e.g., Harlow, 2015 and Bauer, 1997), it is considered to be low-resourced in the context of NLP. An overview of Māori corpora is given in Chapter 6.

Māori is simultaneously threatened (endangered) and undergoing revitalisation (Bell et al., 2005). Recent statistics show that Māori is spoken by 3% of the New Zealand population (Te Tari Mātāwaka, 2020), including roughly one in six Māori adults (Te Kupenga, 2020). All of these adults are bilingual English speakers, with most having learned Māori as a second language rather than as their first. Reporting on *Tu Kupenga 2018*, Lane (2024) highlights a critical increase in the number of speakers belonging to the youngest generation (born between 1984–2003). Focusing on the most proficient speakers, he estimates that 22% of Māori adults from the youngest generation, 18% of adults born between 1964–1983, and 23% of older adults (born before 1964) can speak Māori fairly well to very well. Lane's findings also suggest that, while Māori-medium schools are important, language acquisition is greatly enriched by speaking Māori at home and in community contexts.

As noted above, although the number of Māori speakers remains low, attitudes towards the Māori language have improved in recent years. More New Zealanders are showing an interest in learning the language, as evidenced by an increase in te reo Māori course enrolments, particularly among non-Māori (Education Counts, 2023; Berardi-Wiltshire and Bortolotto, 2022). Meanwhile, Māori language use is becoming more prominent in English-medium schools (May, 2023). An attitudes survey from 2021 indicated that the majority (62%) of New Zealanders are in favour of Māori being taught at primary

[6] See `https://glottolog.org/resource/languoid/id/maor1246`

school (Ruru, 2022), with 57% also agreeing that the government should promote the use of Māori in everyday contexts.

2.3.1.1 Social Media Presence

Anecdotally, there is a growing community of Māori-language speakers who use Māori on social media, including Facebook, Twitter/X, Reddit, TikTok and Instagram (Keegan, 2019; Keegan et al., 2015; Trye et al., 2019). This is believed to increase the prestige and vitality of the language (Keegan and Cunliffe, 2014). Social media play a significant role in our everyday lives and can provide rich insights into expressions of identity and ideology. Furthermore, given the vast amount of data available online, this genre offers unique opportunities for linguistic analysis. However, as far as we are aware, no quantitative analyses of any aspect of Māori language on social media have previously been carried out. We address this gap in the existing body of knowledge through a case study of possession in Māori-language tweets, which is presented in Chapter 8.

2.3.2 Language Contact in Aotearoa New Zealand

The Māori people, believed to be the first inhabitants of Aotearoa, are thought to have arrived from East Polynesia approximately 800–1,000 years ago (McLauchlan, 2014, p. 27). Their language quickly evolved in response to the local environment, distinguishing it from its Polynesian predecessors. While Māori was the only language spoken in mainland New Zealand before the arrival of European settlers in the late 18th century, eventually the majority of the Indigenous people would go from being monolingual Māori speakers, to bilinguals, to monolingual English speakers (Spolsky, 2005).

Captain James Cook, a British explorer and naval officer, first visited the shores of New Zealand in 1769, bringing Māori into initial contact with the English language. However, significant European immigration did not occur for several more decades. Interactions between British settlers and Indigenous Māori initially displayed elements of a promising partnership, characterised by mutual cooperation and exchange. Along with whalers, sealers and various coastal traders, early Anglican missionaries had arrived in Aotearoa by the 19th century. With a view to setting up a Christian mission station in the land, most missionaries became fluent in te reo Māori. They were also the first to devise a written form of the language, with Samuel Marsden compiling the first Māori vocabulary list and Thomas Kendall publishing the first Māori

language dictionary in 1815 (McLauchlan, 2014, pp. 54–55; Whaanga and Greensill, 2014, pp. 16–20).

The partnership between the Māori people and British settlers was formalised in 1840 with *Te Tiriti o Waitangi* ('The Treaty of Waitangi'), which profoundly changed New Zealand's demographic, political and social fabric (Belich, 2002). Resulting from a meeting held in Waitangi, this document was signed by representatives of the British Government and leaders from most Māori tribes. The Treaty established New Zealand as a Crown Colony and appointed Captain William Hobson as Governor. It ostensibly guaranteed Māori control and ownership over their lands and resources, both material and cultural, including the Māori language (May and Hill, 2018, p. 309).

Contrary to these assurances, the British Government soon began to exert dominance through legislation that was detrimental to the Indigenous people. During and after the New Zealand Wars (1845–1872), Māori lost their land through government-imposed confiscations and faced devastating casualties from war and disease. Over the ensuing decades, they were were subjected to laws that "undermined their self-determination, leading to political marginalisation, the alienation of Māori land, and intergenerational impoverishment and racism" (Te Kāhui Tika Tangata, 2022, p. 45).

Following sustained attempts to assimilate Māori into the growing Pākehā[7] population, the Māori language would soon also become severely threatened. Early mission schools were initially Māori-medium institutions in which bilingual education was valued and successful, but were switched to an English-medium system from the 1840s. Under the Native Schools Act 1867, Māori were physically punished for speaking their language (Whaanga and Greensill, 2014, p. 9; King, 2018, p. 593), which contributed to a break in intergenerational transmission (Fishman, 1991). Furthermore, urbanisation of Māori after World War II saw the Māori population shift from being 90% rural to 80% urban in less than two decades. This was accompanied by a severe reduction in the amount of Māori spoken at home and within local communities (May, 2023). All of this precipitated the systematic delegitimisation of te reo Māori (Benton, 1988), steering it towards a pathway of decline and ultimately jeopardising its future survival. As May (2023, p. 665) points out, "It is this context of rapid language loss that galvanized the Māori language revitalization movement from the early 1980s onward."

[7]The term *Pākehā* refers to New Zealanders of European descent.

2.3.3 Māori-Language Revitalisation

Māori-language revitalisation can be seen as an ongoing negotiation between Māori and Pākehā. Māori-led revitalisation initiatives began in earnest in the 1980s, focusing particularly on the schooling system. However, earlier campaigns by the Māori people set the scene for this to happen. On 14 September 1972, Hana Te Hamara and other Māori activists descended on Parliament to deliver a petition seeking the inclusion of te reo Māori in the New Zealand education system. In recognition of the Crown's failure to fulfil its Treaty obligations, the Waitangi Tribunal was established in 1975. Past and present breaches of the Treaty were subsequently lodged, including a claim relating to inadequate protection of the Māori language, led by activist Te Huirangi Waikerepuru. As a result of this movement, the Māori Language Act was passed in 1987, under which te reo Māori was made an official language of Aotearoa and Te Taura Whiri i te Reo Māori ('The Māori Language Commission') was established. That same year, the first Māori immersion (pre)schools, known as *kōhanga reo* ('language nests'), were also opened (King, 2001). Subsequently, Māori-language immersion primary schools (*kura kaupapa Māori*), secondary schools (*wharekura*) and tertiary institutions (*Wānanga Māori*) were established. These initiatives have earned Māori an international reputation as a successful example of Indigenous language education (May and Hill, 2018). Other significant developments have included support for 21 Māori radio stations and a government-funded Māori television station that began in 2004.

Despite sustained efforts towards revitalisation of te reo Māori, current strategies remain hampered by three major factors: (i) a lack of fluent Māori teachers; (ii) an increasingly aged population of speakers; and (iii) a lack of Indigenous *tino rangatiratanga* ('self-governance') over Māori-language immersion education, which has changed from being a Māori-led and funded initiative to falling under the remit of the NZ Ministry of Education (May and Hill, 2018, pp. 310-311).

Presently, the New Zealand Government aims to ensure that, by 2040, one million New Zealanders will have at least basic proficiency in Māori, with 85% of the population viewing the language as an intrinsic part of their national identity (Te Puni Kōkiri, 2019). This reflects a growing emphasis on non-Maori learning Māori, and on encouraging all New Zealanders to embrace speaking the language, regardless of their level of experience.

2.3.4 New Zealand English (NZE)

English is the dominant language of New Zealand, spoken by roughly 95% of the population (Statistics New Zealand, 2018). Despite current aspirations, New Zealand has been described as an "unusually monolingual country" (Bell and Kuiper, 1999, p. 13) owing to its geographical isolation and settler-colonial history. While NZE is a relatively recent variety, it has acquired "local prestige and is now something that many younger New Zealanders claim as part of their identity" (Bell et al., 2005, p. 13). NZE shares many features with the other Southern Hemisphere Englishes, namely Australian and South African English, but also has its own special characteristics.

Since the 1980s, a large body of work on NZE has emerged, mostly concerning pronunciation (i.e., the New Zealand accent). The English spoken by children in Aotearoa was recognised as distinctive by the early 1900s (Hay et al., 2008). Key features of NZE include the centralised pronunciation of the vowel in words such as 'kit', 'fish' and 'chips' (Bauer, 1994; Bell, 1997), the near merger of words with the phoneme pairs 'ear/air' (Holmes and Bell, 1992; Batterham, 2000), the use of high-rising terminal intonation (Warren, 2005) and rhythm (Nokes and Hay, 2012).

A variety of research on the grammar and morpho-syntax of NZE has also been carried out (e.g., Bauer, 2007; Hundt, 1998, 2008), as well as lexical and pragmatic features, such as the adverb *heaps* (Calude, 2019) and the particle *eh* (Meyerhoff, 1994; Schweinberger, 2018). However, the most unique aspect of NZE concerns the use of Māori words, as detailed below.

2.3.4.1 Māori Loanwords

Loanwords (or borrowings)[8] are commonly cited as the most distinctive feature of NZE (Deverson, 1991; Macalister, 2004, 2006; Hay et al., 2008). These words arise when lexical material is transferred from a source language to a receptor language (Zenner and Kristiansen, 2013, p. 1). In the case of NZE, the source language is Māori (an endangered Indigenous language) and the receptor language is English (a dominant *lingua franca*), which is a highly unusual direction of lexical transfer (Trye et al., 2020). Borrowings also occur in the opposite direction, from English to Māori, but they are not the focus here.

Macalister (2006, p. 18) proposed two main 'waves' of borrowing from Māori: an initial wave during the 'colonisation phase', from the time of first

[8]We use the terms 'loanword' and 'borrowing' interchangeably throughout the thesis.

European arrival in Aotearoa until 1880, and a second wave during the so-called 'decolonisation phase', from 1970 onwards. The intervening period of recolonisation, 1880-1970, saw a resistance to borrowing. Interestingly, each wave was associated with different types of borrowings: the first wave included flora and fauna terms and proper nouns (e.g., *kumara* 'sweet potato', *Hēmi* 'James'), while the second wave had a higher concentration of social and material loanwords (e.g., *kaitiakitanga* 'guardianship', *rohe* 'tribal boundary'). The second wave is strongly linked to te reo Māori revitalisation efforts, and increased prestige of Māori language and culture in general.

Today, Māori loanwords are increasingly used in spoken and written discourse, by monolingual and bilingual New Zealanders, both within and beyond the Māori community (Macalister, 2006; Calude et al., 2020a). There are close to 1,000 borrowings listed in *A Dictionary of Maori Words in New Zealand English* (Macalister, 2005). However, the use of Māori loanwords in NZE is complex, and may vary according to the identity of the author or speaker, as well as the genre and topic of the text. For example, loanwords are prolific in education and schooling domains, in certain media (like Radio New Zealand), as well as in discourse related to Māori language or culture (Calude et al., 2019; Degani et al., 2010). In terms of speaker identity, Māori women use the highest proportion of borrowings (Calude et al., 2020a). Words also differ with respect to how entrenched they are: some words, like 'Māori' itself, are so familiar that speakers do not register their origin as Māori.

It is estimated that the average NZE speaker has passive knowledge of at least 70 to 80 Māori loanwords (Macalister, 2004). Recent studies have shown that non-Māori-speaking New Zealanders acquire considerable *subconscious* knowledge of Māori through ambient exposure to the Māori language (Oh et al., 2020; Panther et al., 2023), but their explicit semantic knowledge is much smaller than their implicit, form-based (proto-lexical) knowledge (Oh et al., 2023).

Māori loanwords have been studied comprehensively across a range of genres over the years (Deverson, 1991; Kennedy and Yamazaki, 2000; De Bres, 2006; Macalister, 2009; Daly, 2016; Calude et al., 2020b,a; Trye et al., 2020). These studies show widespread, productive and ongoing use of words of Māori origin in NZE. In Chapter 9, we will take a novel approach by investigating whether multiple loanword types are used within the same texts, rather than considering only their individual raw frequencies.

2.4 Postscript

This chapter has established a foundation for understanding what follows in this thesis. We began by introducing key concepts in Information Visualisation, focusing on the Visualisation Pipeline, representation, interaction and multidimensional data. We then outlined current visualisation approaches in corpus linguistics, before explaining why this domain is particularly well suited for exploring multidimensional categorical data. Finally, the chapter provided background information about te reo Māori and New Zealand English, including an historical overview of the contact between the Indigenous population and European settlers, which has profoundly shaped the linguistic landscape of Aotearoa today.

Looking ahead, Part II of the thesis will examine visualisation methods for representing multidimensional categorical data, including both new and existing techniques. Part III will introduce new datasets and then apply these visualisation techniques, among others, to case studies concerning Māori and New Zealand English.

2.5 References

Albury, N. J. (2015). Your language or ours? Inclusion and exclusion of non-indigenous majorities in Māori and Sámi language revitalization policy. *Current Issues in Language Planning*, 16(3):315–334.

Albury, N. J. (2016). Defining Māori language revitalisation: A project in folk linguistics. *Journal of Sociolinguistics*, 20(3):287–311.

Allen, W. (2017). Making corpus data visible: Visualising text with research intermediaries. *Corpora*, 12(3):459–482.

Almutairi, B. A. A. (2013). Visualizing patterns of appraisal in texts and corpora. *Text & Talk*, 33(4-5):691–723.

Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE.

Ankerst, M., Keim, D. A., and Kriegel, H.-P. (1996). Circle segments : A technique for visually exploring large multidimensional data sets. In *Visualization '96, Hot Topic Session, San Francisco, CA, November, 1996*.

Anthony, L. (2018). Visualisation in corpus-based discourse studies. In *Corpus Approaches to Discourse*, pages 197–224. Routledge.

Anthony, L. (2023). Antconc (4.2.4). `https://www.laurenceanthony.net/software/antconc/`.

Batterham, M. A. (2000). The apparent merger of the front centring diphthongs—EAR and AIR—in New Zealand English. In Bell, A. and Kuiper, K., editors, *New Zealand English*, pages 111–145. Victoria University Press, Wellington.

Bauer, L. (1994). English in New Zealand. the Cambridge history of the English language: Volume V.

Bauer, L. (2007). Some grammatical features of New Zealand English. *New Zealand English Journal*, 21:1–25.

Bauer, W., Parker, W., Evans, T., and Teepa, T. (1997). *The Reed Reference Grammar of Māori*. Reed.

Becker, R. A. and Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, 29(2):127–142.

Bederson, B. B. and Hollan, J. D. (1994). Pad++: A zooming graphical interface for exploring alternate interface physics. In *Proceedings of the 7th annual ACM symposium on User interface software and technology*, pages 17–26.

Belich, J. (2002). *Paradise Reforged: A History of the New Zealanders from the 1880s to the Year 2000*. University of Hawaii Press.

Bell, A. (1997). The phonetics of fish and chips in New Zealand: Marking national and ethnic identities. *English World-Wide*, 18(2):243–270.

Bell, A., Harlow, R., and Starks, D. (2005). *Languages of New Zealand*. Victoria University Press.

Bell, A. and Kuiper, K. (1999). *New Zealand English*, volume 25. John Benjamins Publishing.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press, Princeton.

Benton, R. A. (1988). The Maori language in New Zealand education. *Language, Culture and Curriculum*, 1(2):75–83.

Berardi-Wiltshire, A. and Bortolotto, M. (2022). Learning Māori in the workplace: Non-Māori learners' assessment of the value of te reo. *International Journal of Bilingual Education and Bilingualism*, 25(9):3463–3474.

Bertin, J. (1983). *Semiology of graphics*. University of Wisconsin Press.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Biggs, B. (1968). The Māori language past and present. In Schwimmer, E., editor, *The Māori People in the Nineteen-Sixties*. Blackwood and Janet Paul, Auckland.

Brath, R. (2020). *Visualizing with text*. CRC Press.

Brehmer, M. and Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics*, 19(12):2376–2385.

Brezina, V. and Platt, W. (2024). #LancsBox X. `http://lancsbox.lancs.ac.uk`.

Burnette, J. and Calude, A. S. (2022). Wake up New Zealand! Directives, politeness and stance in Twitter #Covid19NZ posts. *Journal of Pragmatics*, 196:6–23.

Butt, M., Hautli-Janisz, A., and Lyding, V. (2020). *LingVis : Visual Analytics for Linguistics*. Number 220 in CSLI lecture notes. CSLI Publications, Stanford, California.

Calude, A. S. (2019). The use of heaps as quantifier and intensifier in New Zealand English. *English Language & Linguistics*, 23(3):531–556.

Calude, A. S. (2023). *The Linguistics of Social Media: An Introduction*. Taylor & Francis.

Calude, A. S. and Delahunty, G. (2020). Just because. Constructions in spoken and written New Zealand English.

Calude, A. S., Miller, S., Harper, S., and Whaanga, H. (2019). Detecting language change: Māori loanwords in a diachronic topic-constrained corpus of New Zealand English newspapers. *Asia and Pacific Variation Journal*, 5(2):109–137.

Calude, A. S., Miller, S., and Pagel, M. (2020a). Modelling loanword success–a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, 16(1):29–66.

Calude, A. S., Stevenson, L., Whaanga, H., and Keegan, T. T. (2020b). The use of Māori words in National Science Challenge online discourse. *Journal of the Royal Society of New Zealand*, 50(4):491–508.

Card, S. K., Mackinlay, J., and Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

Carpendale, M. S. T. (2003). Considering visual variables as a basis for information visualisation. *Computer Science TR #2001-693*, 16.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368.

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American statistical association*, 79(387):531–554.

Cockburn, A., Karlson, A., and Bederson, B. B. (2009). A review of overview+

detail, zooming, and focus+ context interfaces. *ACM Computing Surveys (CSUR)*, 41(1):1–31.

Coekin, J. (1969). A versatile presentation of parameters for rapid recognition of total state. In *Proceedings International Symposium on Man-Machine Systems, IEEE Conference Record*, volume 69.

Collins, C. M. (2010). *Interactive visualizations of natural language*. University of Toronto.

Culy, C. and Lyding, V. (2010). Visualizations for exploratory corpus and text analysis. In *Proceedings of the 2nd International Conference on Corpus Linguistics (CILC-10)*, pages 257–268, A Coruña.

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

Davies, M. (2020). English-corpora.org: A guided tour. Available online. `https://www.english-corpora.org/pdf/english-corpora.pdf`.

De Bres, J. (2006). Maori lexical items in the mainstream television news in New Zealand. *New Zealand English Journal*, 20:17–34.

Degani, M. et al. (2010). The Pakeha myth of one New Zealand/Aotearoa. An exploration in the use of Maori loanwords in New Zealand English. In *From international to local English-and back again*, pages 165–196. Peter Lang.

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

Education Counts (2023). Tertiary enrolments in te reo Māori language courses, and other languages. `https://www.educationcounts.govt.nz/statistics/tertiary-enrolments-in-language-courses,-including-te-reo-maori-courses`.

Fishman, J. A. (1991). *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*, volume 76. Multilingual matters.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., and Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the public interest*, 22(3):110–161.

Harlow, R. (2007). *Maori: A Linguistic Introduction*. Cambridge University Press.

Harlow, R. (2015). *A Māori Reference Grammar (2nd ed.)*. Huia Publishers.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In *Computer science and statistics: Proceedings of the 13th symposium on the interface*, pages 268–273. Springer.

Hashimoto, D. (2019). *Loanword phonology in New Zealand English: Exemplar*

activation and message predictability. PhD thesis, University of Canterbury.

Hay, J., Maclagan, M., and Gordon, E. (2008). Dialects of English. In *New Zealand English*. Edinburgh University Press, Edinburgh, UK.

Healey, C. G., Booth, K. S., and Enns, J. T. (1996). High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(2):107–135.

Heer, J. and Shneiderman, B. (2012). Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue*, 10(2):30–55.

Hilpert, M. (2011). Dynamic visualizations of language change: Motion charts on the basis of bivariate and multivariate data from diachronic corpora. *International Journal of Corpus Linguistics*, 16(4):435–461.

Holmes, J. and Bell, A. (1992). On shear markets and sharing sheep: The merger of EAR and AIR diphthongs in New Zealand English. *Language Variation and Change*, 4(3):251–273.

Hundt, M. (1998). New Zealand English grammar: Fact or fiction? *New Zealand English Grammar*, pages 1–228.

Hundt, M., Hay, J., and Gordon, E. (2008). New Zealand English: Morphosyntax. In Burridge, K. and Kortmann, B., editors, *Varieties of English 3: The Pacific and Australasia*, pages 305–340. Mouton de Gruyter, Berlin.

Inselberg, A. (1985). The plane with parallel coordinates. *The visual computer*, 1:69–91.

Isaacs, L., Odlum, A., and León-Araúz, P. (2024). Quartz: A template for quantitative corpus data visualization tools. *Languages*, 9(3):81.

Jänicke, S., Franzini, G., Cheema, M. F., and Scheuermann, G. (2015). On close and distant reading in digital humanities: A survey and future challenges. *EuroVis (STARs)*, 2015:83–103.

Keegan, T. T. (2019). Māori language procreation on social media. In *Proceedings of the 17th International Conference on Minority Languages (ICML XVII)*, pages 31–31, Leeuwarden, NL.

Keegan, T. T. and Cunliffe, D. (2014). Young people, technology and the future of te reo Māori. In Higgins, R., Rewi, P., and Olsen-Reeder, V., editors, *The value of the Māori language: Te Hua o te Reo Māori*, pages 385–398. Huia Publishers.

Keegan, T. T., Mato, P., and Ruru, S. (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.

Keim, D. (2000). Designing pixel-oriented visualization techniques: Theory

and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78.

Keim, D. A., Kriegel, H.-P., and Ankerst, M. (1995). Recursive pattern: A technique for visualizing very large amounts of data. In *Proceedings Visualization'95*, pages 279–286. IEEE.

Kennedy, G. and Yamazaki, S. (2000). The influence of Maori on the New Zealand English lexicon. In *Corpora Galore*, pages 33–44. Brill.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1:7–36.

King, J. (2001). Te kohanga reo: Maori language revitalization. In Hinton, L. and Hale, K., editors, *The Green Book of Language Revitalization in Practice*, pages 119–131. Academic Press, New York.

King, J. (2018). Māori: Revitalization of an endangered language. In Rehg, K. L. and Campbell, L., editors, *The Oxford handbook of endangered languages*, pages 592–612. Oxford University Press.

Koffka, K. (1935). *Principles of Gestalt Psychology*. Lund Humphries.

Kucher, K. and Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific visualization symposium (pacificVis)*, pages 117–121. IEEE.

Lane, C. (2024). First and second language speakers in the revitalisation of te reo Māori: A statistical analysis from Te Kupenga 2018. *Te Reo*, 66(2):28–56.

LeBlanc, J., Ward, M. O., and Wittels, N. (1990). Exploring n-dimensional databases. In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pages 230–237. IEEE.

Liu, S., Maljovec, D., Wang, B., Bremer, P.-T., and Pascucci, V. (2016). Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, 23(3):1249–1268.

Liu, Z., Navathe, S. B., and Stasko, J. T. (2011). Network-based visual analysis of tabular data. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 41–50. IEEE.

Macalister, J. (2004). Listening to proper nouns: Social change and Maori proper noun use in New Zealand English. *New Zealand English Journal*, 18:24–34.

Macalister, J., editor (2005). *A Dictionary of Maori Words in New Zealand English*. Oxford University Press.

Macalister, J. (2006). The Maori presence in the New Zealand English lexi-

con, 1850–2000: Evidence from a corpus-based study. *English World-Wide*, 27(1):1–24.

Macalister, J. (2009). Investigating the changing use of te reo. *NZ Words*, 13(1):3–4.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions On Graphics (Tog)*, 5(2):110–141.

Marras Tate, J. and Rapatahana, V. (2022). Māori ways of speaking: Code-switching in parliamentary discourse, Māori and river identity, and the power of Kaitiakitanga for conservation. *Journal of International and Intercultural Communication*, pages 1–22.

May, S. (2023). New Zealand is "racist as f**k": Linguistic racism and te reo Māori. *Ethnicities*, 23(5):662–679.

May, S. and Hill, R. (2018). Language revitalization in Aotearoa/New Zealand. In Hinton, L., Huss, L., and Roche, G., editors, *The Routledge handbook of language revitalization*, pages 309–319. Routledge.

McEnery, T. and Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McLauchlan, G. (2014). *A Short History of New Zealand*. Penguin.

Meyerhoff, M. (1994). Sounds pretty ethnic, eh?: A pragmatic particle in new zealand english. *Language in Society*, 23(3):367–388.

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Nokes, J. and Hay, J. (2012). Acoustic correlates of rhythm in New Zealand English: A diachronic study. *Language Variation and Change*, 24(1):1–31.

Oh, Y., Todd, S., Beckner, C., Hay, J., King, J., and Needle, J. (2020). Non-Māori-speaking New Zealanders have a Māori proto-lexicon. *Scientific reports*, 10(1):22318.

Oh, Y. M., Todd, S., Beckner, C., Hay, J., and King, J. (2023). Assessing the size of non-Māori-speakers' active Māori lexicon. *Plos one*, 18(8).

Panther, F. A., Mattingley, W., Todd, S., Hay, J., and King, J. (2023). Proto-lexicon size and phonotactic knowledge are linked in non-Māori speaking New Zealand adults. *Laboratory Phonology*, 14(1).

Rao, R. and Card, S. K. (1994). The Table Lens: Merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322.

Rayson, P., Mariani, J., Anderson-Cooper, B., Baron, A., Gullick, D., Moore, A., and Wattam, S. (2016). Towards interactive multidimensional visualisa-

tions for corpus linguistics. *Journal for language technology and computational linguistics*, 31(1):27–49.

Roberts, J. C. (2007). State of the art: Coordinated & multiple views in exploratory visualization. In *Fifth international conference on coordinated and multiple views in exploratory visualization (CMV 2007)*, pages 61–71. IEEE.

Ruru, K. (2022). Māori should be a core subject in primary schools. Stuff NZ. `https://www.stuff.co.nz/pou-tiaki/te-reo-maori/300629716/three-in-five-kiwis-think-te-reomori-should-be-a-core-subject/-in-primary-schools`.

Schneider, G., El-Assady, M., and Lehmann, H. M. (2017). Tools and methods for processing and visualizing large corpora. *Studies in Variation, Contacts and Change in English*, 19.

Schweinberger, M. (2018). The discourse particle eh in New Zealand English. *Australian Journal of Linguistics*, 38(3):395–420.

Scott, M. (2024). WordSmith Tools, version 9 (64 bit version).

Seyser, D. and Zeiller, M. (2018). Scrollytelling–an analysis of visual storytelling in online journalism. In *2018 22nd international conference information visualisation (IV)*, pages 401–406. IEEE.

Sheehan, S., Masoodian, M., and Luz, S. (2022). Task-based quantitative evaluation of the Concordance Mosaic visualization. In *2022 26th International Conference Information Visualisation (IV)*, pages 123–129. IEEE.

Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE software*, 11(6):70–77.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pages 336–343. IEEE.

Siirtola, H., Isokoski, P., Säily, T., and Nevalainen, T. (2016). Interactive text visualization with Text Variation Explorer. In *2016 20th International Conference Information Visualisation (IV)*, pages 330–335. IEEE.

Siirtola, H., Nevalainen, T., Säily, T., and Räihä, K.-J. (2011). Visualisation of text corpora: A case study of the PCEEC. In *How to deal with data: Problems and approaches to the investigation of the English language over time and space*. Varieng.

Siirtola, H., Säily, T., Nevalainen, T., and Räihä, K.-J. (2014). Text variation explorer: Towards interactive visualization tools for corpus linguistics. *International Journal of Corpus Linguistics*, 19(3):417–429.

Spolsky, B. (2005). Māori lost and regained. *Languages of New Zealand*,

page 67.

Statistics New Zealand (2018). 2018 census totals by topic: National highlights updated. https://www.stats.govt.nz/information-releases/2018-census-totals-by-topic-national-highlights-updated.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

Stolte, C., Tang, D., and Hanrahan, P. (2008). Polaris: A system for query, analysis, and visualization of multidimensional databases. *Communications of the ACM*, 51(11):75–84.

Sönning, L. and Schützler, O. (2023). Data visualization in corpus linguistics: Critical reflections and future directions. In Sönning, L. and Schützler, O., editors, *Data Visualization in Corpus Linguistics: Critical Reflections and Future Directions*, number 22 in Studies in Variation, Contacts and Change in English. VARIENG, Helsinki.

Te Kupenga (2020). More than 1 in 6 Māori people speak te reo Māori. https://www.stats.govt.nz/news/more-than-1-in-6-maori-people-speak-te-reo-maori.

Te Kāhui Tika Tangata (2022). Maranga Mai! The dynamics and impacts of white supremacy, racism, and colonisation upon tangata whenua in Aotearoa New Zealand. https://admin.tikatangata.org.nz/assets/Documents/Maranga-Mai_Full-Report_PDF.pdf.

Te Puni Kōkiri (2010). 2009 survey of attitudes, values, and beliefs towards the Māori language.

Te Puni Kōkiri (2019). Maihi Karauna: The Crown's strategy for Māori language revitalisation 2019–2023. https://www.tpk.govt.nz/en/o-matou-mohiotanga/te-reo-maori/crowns-strategy-for-maori-language-revitalisation.

Te Tari Mātāwaka (2020). Languages spoken in New Zealand. www.ethniccommunities.govt.nz.

Theron, R. and Fontanillo, L. (2015). Diachronic-information visualization in historical dictionaries. *Information Visualization*, 14(2):111–136.

Theus, M. (2002). Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7:1–9.

Tominski, C. (2022). *Interaction for visualization*. Springer Nature.

Trye, D., Calude, A., Bravo-Marquez, F., and Keegan, T. T. (2019). Māori loanwords: A corpus of New Zealand English tweets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 136–142, Florence. Association for Com-

putational Linguistics.

Trye, D., Calude, A. S., Bravo-Marquez, F., and Keegan, T. T. (2020). Hybrid hashtags: #Youknowyoureakiwiwhen your tweet contains Māori and English. *Frontiers in Artificial Intelligence*, 3:15.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Warren, P. (2005). Patterns of late rising in New Zealand English: Intonational variation or intonational change? *Language Variation and Change*, 17(2):209–230.

Whaanga, H. and Greensill, H. (2014). An account of the evolution of language description of te reo Maori since first contact. In Onysko, A., Degani, M., and King, J., editors, *He Hiringa, He Pūmanawa: Studies on the Māori language*, pages 7–32. Huia.

Yi, J. S., ah Kang, Y., Stasko, J., and Jacko, J. A. (2007). Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231.

Zenner, E. and Kristiansen, G. (2013). *New perspectives on lexical borrowing: Onomasiological, methodological and phraseological innovations*, volume 7. Walter de Gruyter.

Zhao, J., Collins, C., Chevalier, F., and Balakrishnan, R. (2013). Interactive exploration of implicit and explicit relations in faceted datasets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2080–2089.

Part II

Visualising Categorical Data

Chapter 3

A Review of Categorical Visualisation Techniques

This chapter serves as the main literature review for the thesis, providing a background and framework for our primary research question: What generalisable information visualisation techniques can be developed or adapted to enable the effective analysis of datasets involving multiple categorical variables?

3.1 Introduction

Categorical variables are prevalent in real-world datasets, frequently occurring in domains such as the behavioural and social sciences, public health, biomedical science, education, business and marketing (Agresti, 2012). Examples of categorical data include responses to multiple-choice survey questions (e.g., strongly disagree, disagree, neutral, agree, strongly agree), treatment options assigned to participants in a medical trial (e.g., drug A, drug B, placebo), and the biological class to which different animals belong (mammal, bird, reptile, etc.). Categorical variables are even found in highly quantitative fields, such as industrial quality control, where products are rated based on their adherence to specific standards (Agresti, 2019).

When dealing with categorical data, analysts are typically interested in comparing category frequencies and investigating relationships between categories. Like other data types, the amount of categorical data available is continually growing, increasing the need for efficient analysis methods (Johansson Fernstad, 2011). Surprisingly, despite these demands, visualisation techniques for categorical data have received considerably less attention in the literature compared to those for numeric data (Liu et al., 2016; Friendly,

1998). This is especially true when the need arises to visualise more than three categorical variables simultaneously.

Categorical data visualisation presents several challenges. Firstly, nominal categories do not have an intrinsic order or inherent spatial mapping (Cibulková and Kupková, 2022). Secondly, combinations of categories become increasingly sparse when more variables are added, exemplifying the 'curse of dimensionality' (Hofmann, 2006). Thirdly, variables with a large number of categories may exceed the limits of the visual encoding, or render a visualisation unreadable. Overall, compared to numeric data, categorical visualisation techniques appear to be more sensitive to structural characteristics of the data (Johansson Fernstad, 2011).

This chapter provides a review and taxonomy of categorical visualisation techniques. We begin by defining key terminology (Section 3.2), before detailing the scope of the review and our method for gathering and organising the relevant literature (Section 3.3). The heart of the chapter describes six distinct 'families' of techniques that we have identified, which form the basis of the proposed taxonomy (Section 3.4). We focus on prototypical examples within each family, then introduce nine different types of analysis tasks from a categorical visualisation perspective (Section 3.5). Finally, we compare general strengths and weaknesses of each family and reflect on opportunities for future work (Section 3.6). An interactive repository of the techniques reviewed in this chapter is available at: `https://cat-vis.github.io/`.

3.2 Categorical Data

Categorical data consist of variables that take a fixed set of values, each representing a distinct category or group, such as colour. Due to their unique characteristics, these variables require different analysis methods from numeric data, including specialised visualisation techniques (Friendly and Meyer, 2015). The main advantage of visualising categorical data is the ability to reveal relationships between multiple variables or categories more clearly than tabular or textual representations.

3.2.1 Terminology

A range of terms is used in the literature to refer to categorical data. Our preferred terms within this thesis are emphasised here in bold. Individual **(data) items** may alternatively be called *objects, cases, records, tuples, points, vectors, observations* or *samples*. The properties of each data item are described

by a set of **variables**, where a variable is defined as a characteristic that can vary from one item to another. Variables are sometimes also known as *attributes, features* or *dimensions*. The number of distinct values that a variable can take is its **cardinality**, while the values themselves are variously referred to as **categories**, *levels* or *classes*. We refer to a group of two or more orthogonal categories as a **combination of categories**. Categorical variables with only two possible values are sometimes referred to as *binary* variables (Agresti, 2019; Friendly and Meyer, 2015).

Following Tan et al. (2006), we consider a categorical variable to be either **nominal**, meaning its categories are unordered, or **ordinal**, meaning they have a natural ordering. Examples of nominal variables include 'gender' and 'continent', whereas 'customer satisfaction' and 'education level' are both ordinal variables. We consider it important for a categorical visualisation tool to accommodate both these data types. Additionally, quantitative (numeric) variables can be **binned**, or *discretized*, to form (typically) ordinal variables, though this process results in a loss of precision. Two common binning strategies are to create categories of equal *width* or *frequency* (Dougherty et al., 1995). For example, 'income' and 'age' are often divided into specific ranges.

Data can be *univariate*, *bivariate*, or *multivariate*, depending on whether they comprise one, two, or more than two variables, respectively. We use the terms *multivariate* and *multidimensional* interchangeably. Multivariate categorical data are relatively common: census data may include variables such as gender, education level, religion and marital status; medical records might include disease types, treatment protocols and patient outcomes; retail databases frequently categorise products by type, payment method and customer demographics. Analysing all categorical variables simultaneously can enhance understanding of complex relationships and support informed decision-making.

Statistical models often distinguish between **response** (or *dependent*) variables and **explanatory** (or *independent* variables). The latter are thought to partially explain the value of the former. Often, a dataset contains a single response variable and several explanatory variables (Theus, 2008). For example, in the Titanic and Mushroom datasets introduced below, the response variables are *Survived* (yes/no) and *Edibility* (poisonous/edible), respectively. Depending on a user's analysis task, it may be beneficial to highlight a response variable within a visualisation by assigning it a prominent position, for instance, or mapping it to colour.

3.2.1.1 Common Datasets and Data Forms

The *Titanic dataset* (Dawson, 1995; see Figure 3.1) is arguably the most well-known dataset in the field of categorical visualisation. This dataset provides socio-historical information about the passengers and crew aboard the RMS *Titanic*, which tragically sank in 1912. Although the dataset has been the subject of considerable attention (see, for example, Symanzik et al., 2019), and is widely used for illustrative purposes, it is relatively small, containing only 4 variables, 10 categories and 2201 data items. Several different versions of the Titanic data exist, some of which include the names of passengers as an additional string-type (text) variable. We will use the Titanic dataset in most of the examples in this chapter.

The synthetic *Mushroom dataset* (Schlimmer, 1987), describing properties of mushrooms like their colour, odour and stalk shape, is considerably larger than the Titanic dataset. It comprises 22 variables, 119 categories and 8124 data items, making it a popular choice for demonstrating how categorical visualisation techniques can (or cannot) scale to larger and more complex datasets.

At the internal representation level, Friendly and Meyer (2015) refer to three main forms of categorical data: *case form*, *frequency form* and *table form*, which are illustrated in Figure 3.1. Case form provides each data item as a separate entry, with rows corresponding to data items and columns to variables. This allows any data item to be traced back to its individual identifier. In contrast, frequency form collapses identical combinations of categories

(a)

ID	Class	Age	Sex	Fate
1	first	adult	male	survived
2	first	adult	male	survived
3	first	adult	male	survived
⋮	⋮	⋮	⋮	⋮
2202	crew	adult	female	died

(b)

Class	Age	Sex	Fate	Freq
crew	adult	male	died	670
third	adult	male	died	387
⋮	⋮	⋮	⋮	⋮
second	child	female	died	0

(c)

		Sex / Fate			
		female		*male*	
Age	Class	*died*	*survived*	*died*	*survived*
adult	*crew*	3	20	670	192
	first	4	140	118	57
	second	13	80	154	14
	third	89	76	387	75
child	*crew*	0	0	0	0
	first	0	1	0	5
	second	0	13	0	11
	third	17	14	35	13

Figure 3.1: The Titanic dataset shown in (a) case form, (b) frequency form and (c) table form. The *Survived* (yes/no) variable from Dawson's (1995) original dataset has been renamed *Fate* (survived/died) to give the two categories semantically descriptive names.

into a single row, reporting their counts in an additional column. Finally, table form presents data in a contingency table, which involves cross-tabulating some or all of the available variables.

3.3 Scope and Methodology

In this review, we focus on visualisation techniques that are capable of showing *purely* categorical data, for any number of variables. We limit our analysis to techniques that treat variables as having *flat* and *disjoint* categories. In other words, the categories within each variable lack *sub*-categories, and are mutually exclusive. Datasets that include multi-value categories are likely better modelled as sets (Alsallakh et al., 2016). Furthermore, our review focuses on exploratory data analysis rather than on statistical model building (see Friendly and Meyer, 2015). Categorical data with special properties fall outside the scope of this review, including geospatial and time-oriented data, as well as relational data with categorical attributes.[1]

This chapter synthesises ideas and techniques for visualising categorical data from roughly 120 papers. The literature was extracted by paying special attention to publications from *IEEE Xplore, EuroGraphics, Sage Information Visualization* and the *Journal of Computational and Graphical Statistics* that explicitly mentioned 'categorical' data in the title or keywords. We also expanded our search to include literature cited by these papers, as well as work that cited them. The collected papers were tagged according to their primary **Contribution**, the vast majority (80%) being *Technique* papers:

- Technique: The paper introduces a specific technique or system for visualising categorical data.
- Evaluation: The paper provides an empirical, algorithmic or theoretical evaluation of visualisation approaches for categorical data.
- Ordering Algorithm: The paper contributes an algorithm for rearranging categorical data.
- Framework: The paper contributes a framework or paradigm for visualising categorical data.
- Textbook: A textbook on the topic of visualising categorical data.
- Survey: The paper presents a survey of categorical data visualisation or a related field.

Technique papers were tagged according to five further attributes that we deemed important, as outlined in Table 3.1.

[1] We do, however, explore this further in our final case study, in Chapter 9.

Table 3.1: Classification system for technique papers.

Category	Description
Family	
Size-Encoding	The technique uses bars (line marks) with the length channel, or wedges (area marks) with the angle or length channels.
Space-Filling	The technique fills the available space and likely imposes a hierarchy of variables.
Table	The technique represents data in a 2D table or matrix, where each cell contains visual encodings.
Glyph	The technique uses glyphs or icons to represent individual items or aggregates in the dataset.
Miscellaneous	The technique represents frequencies (in line with the *CatViz* approach) but does not fit into any of the above categories.
Projection	The technique converts categories into numerical values before representing these visually (in line with the *QuantViz* approach).
Data Type	
Homogeneous	The technique only supports categorical (not quantitative) data.
Heterogeneous	The technique supports a mixture of categorical and quantitative data.
Dimensionality	
Univariate	The technique supports only one categorical variable.
Bivariate	The technique supports up to (or exactly) two categorical variables.
Trivariate	The technique supports up to (or exactly) three categorical variables.
Multivariate	The technique can support more than three categorical variables.
Cardinality	
Very Low	The technique requires at least one binary variable.
Low	The technique supports variables with roughly (only) 2-5 categories.
Moderate	The technique can handle at least one variable with 6-10 categories.
High	The technique is designed to support at least one variable with 10-100 categories.
Very High	The technique is designed to support at least one variable with 100+ categories.
Alignment	
Linear	The technique arranges data along perpendicular or parallel axes.
Radial	The technique is laid out in elliptical form, and likely uses polar coordinates.
Other	The technique does not use a linear or radial layout (e.g., force-directed).

The six 'families', which form the basis of our proposed taxonomy, are explained in detail in Section 3.4. It was sometimes necessary to make subjective judgements when assigning these tags, if relevant details were not overtly mentioned in the paper. We acknowledge that the interplay between a technique's

supported cardinality and dimensionality is important, though this was not explicitly coded. Our final literature collection can be interactively explored at https://cat-vis.github.io, which was created using the *SurVis* template (Beck et al., 2015).

3.3.1 Technique Taxonomy

Given the focus of this thesis on visualisation methods, the technique papers were fundamental to the current review. We have organised this body of literature into a two-level taxonomy, as shown in Figure 3.2. The first-level classification groups techniques into *CatViz* (frequency-based) and *QuantViz* (quantification-based) approaches, following Johansson Fernstad and Johansson (2011).

The CatViz approach involves directly mapping the cell counts from a contingency table, using a visual representation suitable for *categorical* data. In contrast, the QuantViz approach projects categories onto a (typically) two-dimensional plane using quantification methods, and then represents the data visually using any technique designed for *numeric* data. The quantification

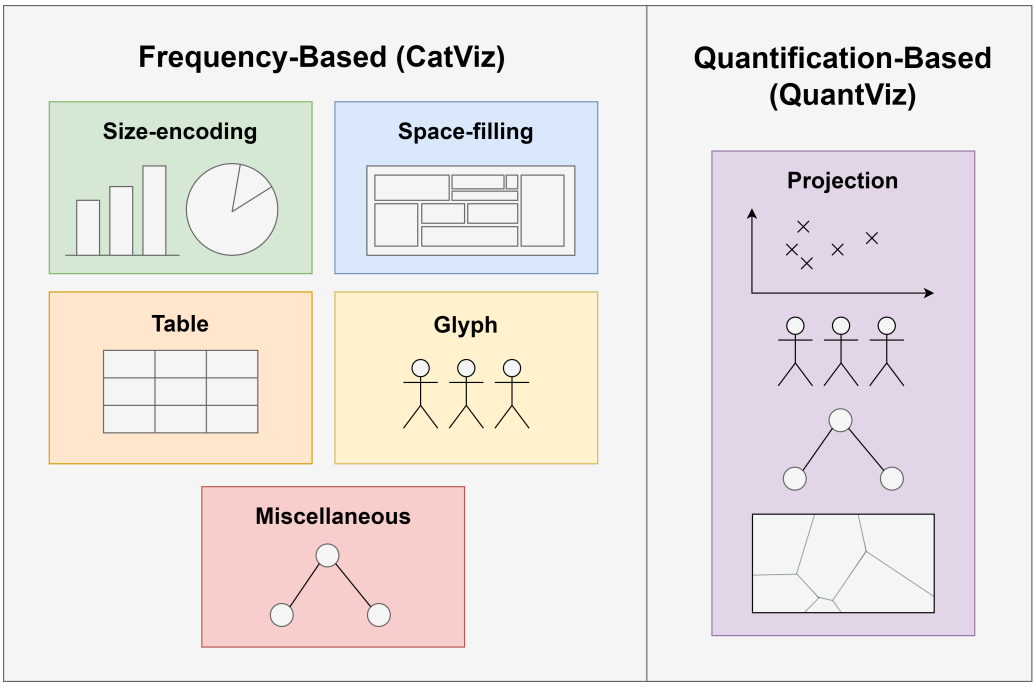


Figure 3.2: Our proposed taxonomy comprises six 'families' of techniques: *size-encoding, space-filling, table, glyph, miscellaneous* (all frequency-based) and *projection* (quantification-based). The rectangle for *projection* is larger to indicate that it encompasses many different possible representations for numeric data.

approach aims to preserve relationships, such as distances, similarities and associations between data points. Each approach has its own merits: in an initial user study (ibid), CatViz techniques were found to be superior for *frequency* tasks (e.g., identifying the most frequent category), while QuantViz techniques were found to be better suited for *similarity* tasks (e.g., determining which two categories are most alike).

In addition, we developed a second-level classification, based on the aforementioned 'families' of visualisation techniques. We have identified six main groups but, as new techniques emerge, others can be added. Five of the six families relate to the CatViz approach: *size-encoding, space-filling, table, glyph* and *miscellaneous*. The remaining category, *projection*, encompasses any visualisation technique used as part of the quantification approach. The projection family is highly versatile, since converting categories to numbers fundamentally changes what can be done with the visual representation. We note that these families are not mutually exclusive: for instance, *dimensional stacking* (Section 3.4.3.2) can be regarded as a hybrid table/space-filling technique.

3.4 Overview of Technique Families

In this section, we describe the six families of techniques, breaking these down into further sub-categories where appropriate. At least one visualisation technique is reviewed in each section, and references are given for related methods.

3.4.1 Size-Encoding Techniques

We define *size-encoding* techniques as those which use *bars* (line marks) with the length channel, or *wedges* (area marks) with the angle or length channels. Consequently, this family can be clearly divided into a bar family and a wedge family. Most techniques in the bar family have *linear* alignment, while those in the wedge family are *radial*. Although equivalent from a mathematical point of view, the wedge family is generally less effective than the bar family, since angles are harder to compare than lengths (Munzner, 2014). The *Trellis display framework* Becker et al. (1996) can be applied to many size-encoding techniques to encode additional categorical data via faceting.

3.4.1.1 Bar Family

Dating back to the latter half of the 18th century (Playfair, 1786, as cited in Friendly, 2006), the *bar chart* (or *column chart*) is a simple yet powerful technique for encoding categorical data. As well as being easy to create and interpret, bar charts are helpful for highlighting precise differences in category counts. For nominal variables, the categories within a bar chart should generally be sorted by frequency (i.e., bar length); for ordinal variables, it may be preferable to preserve the natural ordering of categories. While the classic bar chart is limited to displaying a single categorical variable, numerous variations exist, many of which enable additional variables to be encoded by leveraging colour, texture and/or faceting. These extensions include:

- *Stacked bar charts* and their variants (see Figure 3.3; Indratmo et al., 2018; Streit and Gehlenborg, 2014):
  - *Grouped bar charts* (also called *clustered bar charts*, *dodged bar charts*, *multiple bar charts*, and *multi-series bar charts*)
  - *100% stacked bar charts* (also called *normalised bar charts*)
  - *Layered bar charts*
  - *Diverging stacked bar charts* (also called a *bidirectional bar chart* if the coloured variable is binary)
  - *Inverting stacked bar charts*
  - *Faceted bar charts*
  - *Relative multiples barcharts (rmb plots)*
- *Linked bar charts* (Hummel, 1996), as implemented in tools like *Mondrian* (Theus, 2002) and *High-D* (Brodbeck and Girardin, 2019)
- *Horizon bars* (Lex et al., 2014)
- *Du Bois wrapped bar charts* (Karduni et al., 2020)
- *Pareto charts* (Wilkinson, 2006)
- *Radial bar charts* (Booshehrian et al., 2011)
- *Circular bar charts* (Skau and Kosara, 2016)

Taking one of the most popular examples from this list, the stacked bar chart (Figure 3.3, top left) typically encodes the frequency of two categorical variables, rather than just one. The first variable determines the categories for the bars along the x- or y-axis, as in a regular bar chart, while the second variable is broken down into segments within each bar. These segments are typically distinguished by colour and are consistently ordered across all bars. Stacked bar charts show the marginal distribution of the first variable, and the conditional distribution of the second variable given the first.

Figure 3.3: Six different variations of stacked bar charts (Indratmo et al., 2018).

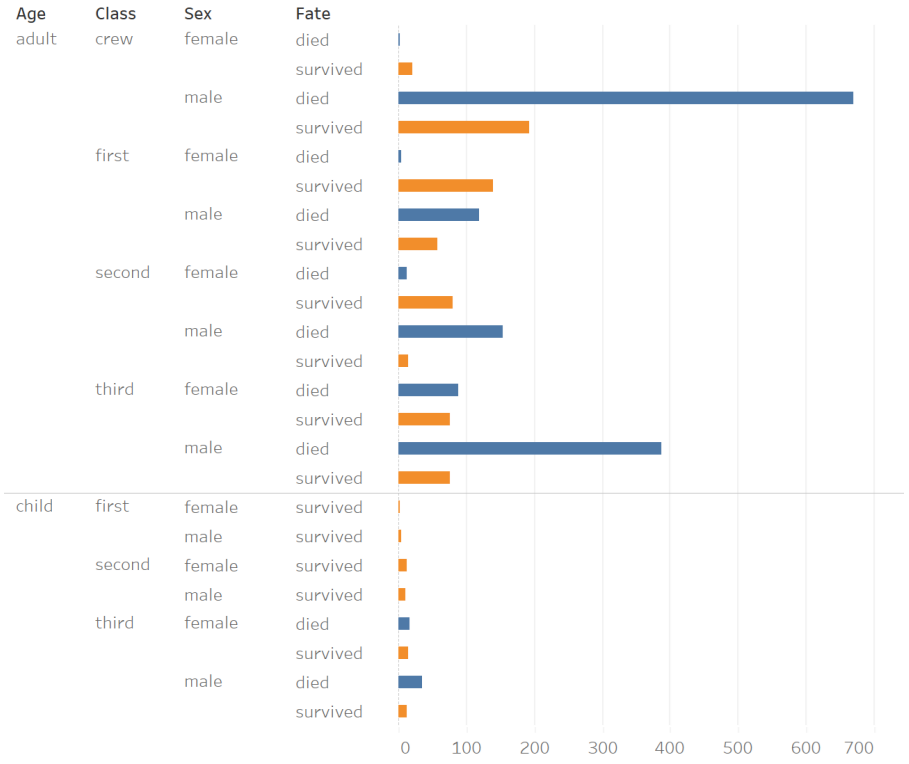


Figure 3.4: 'Multivariate' bar chart showing the joint frequency of all four variables from the Titanic dataset. Colour redundantly encodes Fate (blue = died, orange = survived). Created in Tableau.

This means that reversing the roles of the variables would result in a different plot and potentially yield different insights.

As with a regular bar chart, the scalability of a stacked bar chart ranges from dozens to hundreds of categories for the axis variable, but is limited to roughly a dozen categories for the second variable Munzner (2014). Comparing both the total length and the bottom segment of each bar is straightforward because they share a common baseline, but comparing other segments is more challenging.

Bar charts can display more than two variables by 'chaining' multiple variables along the same or different axes, as shown in Figure 3.4. The bars in the resulting visualisation show the joint frequency of each combination of categories involving all variables. Dozens to hundreds of bars can be shown, and up to roughly eight variables. However, the more variables that are shown, the less room there is to display the labels for each category. This kind of visualisation imposes a hierarchy of variables (like most *space-filling* techniques), which means changing the order of variables can affect the patterns seen, even though the values of the bars remain unchanged. Tooltips and drag-and-drop reordering may help to make sense of patterns in the data.

3.4.1.2 Wedge Family

Members of the wedge family use area marks, rather than line marks, to show frequency. *Pie charts* (Playfair, 1801) and their close cousins, *donut charts* (Skau and Kosara, 2016), are useful for representing proportions or percentages of a whole when there are 12 categories or fewer. They are effective for comparing one category relative to the whole dataset, but not for comparing the proportion of one category to another, except when the variation is extreme, or there are only two categories. Other members of this family include:

- *Nightingale rose chart* (Nightingale, 1857), also known as *sector graphics, Coxcomb charts* and *polar area diagrams*
- *Wind roses* Sanderson and Peacock (2020)
- *Four-fold displays* (Fienberg, 1975; Friendly, 1995))

Although aesthetically pleasing, perceptually, pie and donut charts are known to be less precise than bar charts. Figure 3.5 provides an example of a *faceted pie chart*, representing three of the four variables in the Titanic dataset. However, such charts should be used with caution. In his book *The Visual Display of Quantitative Information*, Edward Tufte (1983, p. 178) remarked: "the only thing worse than a pie chart is several of them, for then the viewer is asked to

Figure 3.5: A faceted pie chart of the Titanic dataset: Class is shown on the x-axis, Sex on the y-axis and Fate is mapped to colour (blue = died, orange = survived). It is clear that many more men than women died in each class.

compare quantities located in spatial disarray both within and between pies". Despite their limitations, pie and donut charts are pervasive and participants in a user study expressed a subjective preference for them over bar charts (Siirtola, 2014).

3.4.2 Space-Filling Techniques

As the name suggests, space-filling techniques are arranged so that the layout consumes all available space in the view. In the context of multivariate categorical data, these techniques typically use area or containment marks to show different combinations of categories. Space-filling techniques are geared towards high information density, but the fact that they use all the available space does not necessarily mean they do so efficiently (Munzner, 2014, p. 175).

A variety of space-filling techniques can be applied to multivariate categorical data by creating a hierarchy of variables (Reza and Watson, 2019; Kosara, 2008). This is despite the fact that the data in question are not inherently hierarchical (i.e., there are no sub-categories). The hierarchy is derived by mapping each categorical variable to a different level, with all categories of the first variable at the top level, all categories of the second variable at the second level, and so on. This results in a fully balanced tree whose nodes represent different combinations of categories. Figure 3.6 shows an example for the Titanic data, together with a corresponding *treemap* (see Section 3.4.2.3).

The order of variables in the hierarchy is significant as it affects the user's ability to perceive structures. This ordering becomes even more crucial as the number of variables increases. It is therefore important for the user to be able

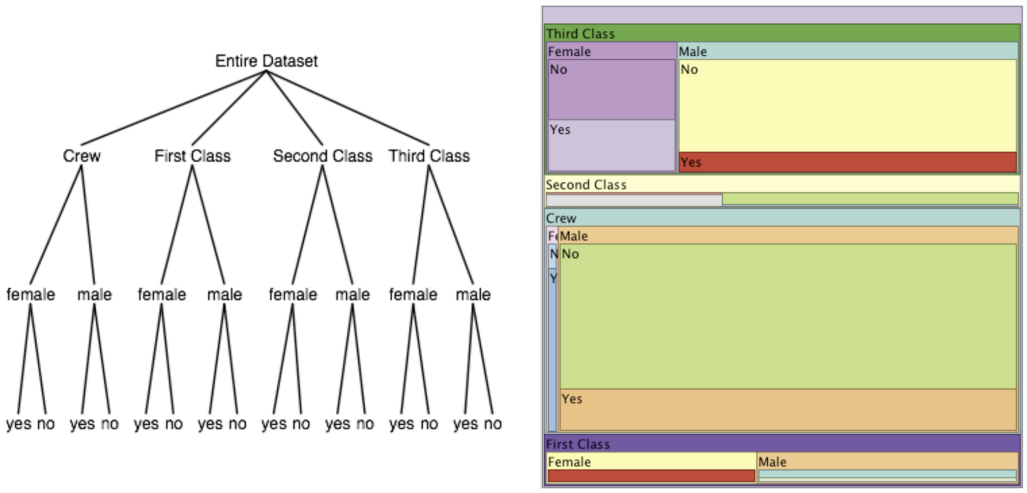


Figure 3.6: Left: Hierarchy derived from three of the four variables in the Titanic dataset, splitting first by Class, then Sex, then Fate. Right: Treemap using the same hierarchical structure, which shows values at the leaves of the tree (the frequency of combinations of all three variables), as well as aggregates at higher levels (Kosara, 2008).

to reorder, add or remove variables as desired (Kosara, 2008). Relevant factors for determining an appropriate order may include the position of the response variable, the perceived importance of other variables, and the distribution of variable cardinalities. An effective ordering for one technique might not work well for another. Colour also plays an important role, and is commonly used to highlight the response variable.

3.4.2.1 ParSets Family

Several categorical visualisation techniques adapt *parallel coordinates* for numeric data (Inselberg, 1985) by substituting data points with a frequency-based representation. *Parallel Sets* (Kosara, 2010; Kosara et al., 2006), pictured in Figure 3.7, is the most well-known technique among this family. Reminiscent of a *Sankey diagram* (Schmidt, 2006), this technique arranges variables along the y-axis in bands of equal width, which are then partitioned according to category frequencies. Associations between subgroups are shown using shaded parallelograms (or ribbons) that connect categories from adjacent dimensions. The widths of individual categories indicate marginal frequencies, while the widths of parallelograms reflect both joint frequencies (relative to the width of the display) and conditional frequencies (relative to the width of the previous subset). Numeric variables can be binned but not shown directly.

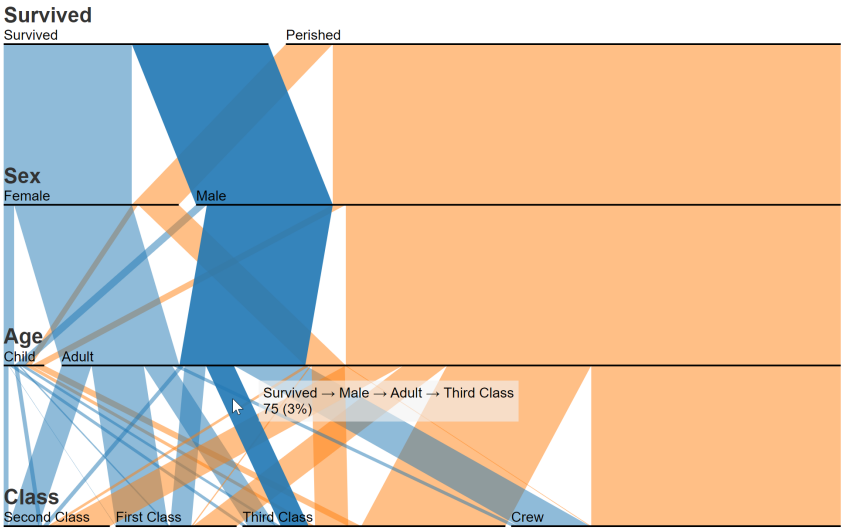


Figure 3.7: A Parallel Sets visualisation of the Titanic dataset, showing all four variables (Davies, 2012).

Two variations of Parallel Sets are possible: *hierarchical* and *pairwise* (see Hofmann and Vendettuoli, 2013). In the hierarchical variation (described above, and shown in Figure 3.7), the parallelograms are split according to every preceding variable, resulting in increasingly complex, and less frequent, subsets. In contrast, the pairwise variation displays two-dimensional subsets relating to each pair of neighbouring variables. The hierarchical view is more useful for visualising multivariate relationships but is inevitably more cluttered.

The main advantage of Parallel Sets is that it can handle roughly 10–15 variables in an interactive environment and 20-30 categories in total, which exceeds the limits of most frequency-based techniques. In addition, the order in which the hierarchy is derived is clearly readable—from top to bottom—and categories and variables can be flexibly reordered, facilitating detection of complex patterns in the data. Parallel Sets can also display numeric variables by binning them.

Key limitations of Parallel Sets include visual interference from line crossings and poor visibility of small parallelograms representing infrequent combinations. These issues are exacerbated when handling large numbers of categories and variables. For example, the Mushroom dataset requires 22 layers and 8123 combinations, which is untenable (Dennig et al., 2024). To alleviate visual clutter, research has focused on measuring and improving the layouts of Parallel Sets (Alsakran et al., 2014; Dennig et al., 2021; Zhang et al., 2019).

Other techniques in the ParSets family, all of which can display mixed data, are *Hammock Plots* (Schonlau, 2003, 2024), *CPCP* (Pilhöfer and Unwin, 2013),

GPCP (VanderPlas et al., 2023), *Parallel Assemblies Plots* (Cantu et al., 2023) and *SET-STAT-MAP* (Wang et al., 2022). Hofmann and Vendettuoli (2013) observed that Parallel Sets and Hammock Plots suffer from the *line width illusion* and *reverse line width illusion*, respectively. They proposed *Common Angle Plots* to overcome these distortions, while Schonlau (2024) suggested a correction to Hammock Plots by replacing the parallelograms with rectangles.

Finally, we note that *chord diagrams* (inspired by Krzywinski, 2009) can be used to visualise relationships between two categorical variables (Humayoun et al., 2018). Chord diagrams are related to techniques in the ParSets family as they emphasise the flow of category subsets, but they are limited to showing only two variables in the same plot.

3.4.2.2 Mosaic Family

Techniques in the mosaic family are largely area-proportional, with colour often being used to highlight particular variables or statistical information. The technique after which this family is named, the *mosaic plot*, was introduced by Hartigan and Kleiner (1981) and further developed by Friendly (1999). An example of a mosaic plot is given in Figure 3.8. In this technique, area-proportional tiles are created by recursively subdividing the space along the axes based on the categories of each variable. In addition to showing joint frequencies through the size of the tiles, mosaic plots show the marginal proportion of the first variable used for splitting, and the conditional proportions for each subsequent variable based on the previous ones. A useful property of mosaic plots is that the cells are aligned when variables are independent Friendly (1999). Unfortunately, mosaic plots become difficult to read when representing more than three variables, or a large number of categories.

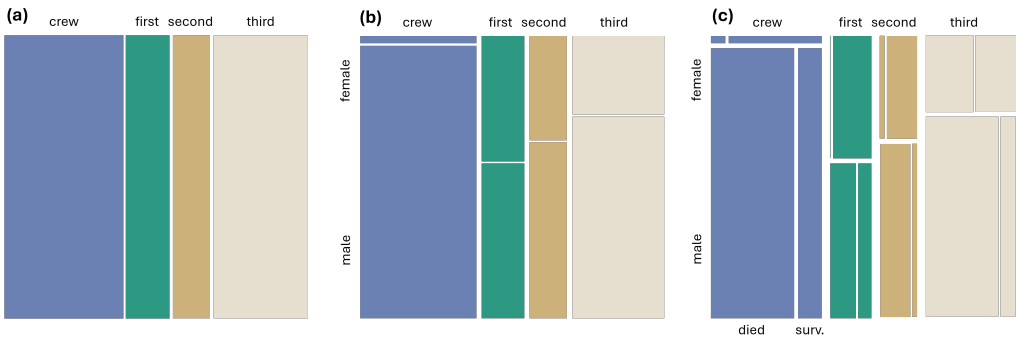


Figure 3.8: Mosaic plot of the Titanic dataset, illustrating the splitting process for three variables: (a) first by Class, (b) then by Sex, (c) then by Survived. Age is not shown. Inspired by Tominski and Schumann (2020).

Residual-based shading of the tiles in a mosaic plot can visually indicate the lack of fit of a specific log-linear model (Friendly, 1994) or the statistical significance of test results (Zeileis et al., 2007). Commonly, two shades for both positive (blue) and negative (red) residuals are used. The shading usually either reflects significance at 90% or 99% confidence levels, or employs fixed cut-offs at ± 2 and ± 4, corresponding to *individual* significance at alpha levels of $\alpha = 0.05$ and $\alpha = 0.001$, respectively (Friendly, 1994). The use of residuals works well for large tiles but not for smaller ones as it is difficult to make out the colours. Moreover, the difference of size and colour may lead to misinterpretations of the data; for instance, if two tiles have the same colour but are drastically different sizes, a viewer may mistakenly believe the larger one has a larger residual.

In addition to the traditional mosaic plot, the mosaic family comprises the following chart types:

- *Spine plots* (Hummel, 1996, Figure 3.8a)
- *Line mosaic plots* (Huh, 2004)
- *Marimekko charts* (Miyamoto et al., 2022)
- *Eikosograms* (Cherry and Oldford, 2003)
- *Double-decker plots* (Hofmann et al., 2000; Hofmann, 2001)
- *Sieve plots* or *parquet diagrams* (Riedwyl and Schüpbach, 1994)
- *Association plots* (Cohen, 1980)
- *Fluctuation diagrams* (Hofmann et al., 2000)
- *Equal bin size plots* (Hofmann et al., 2000)
- *Faceted mosaic plots* (Meyer et al., 2008)
- Further variations arising from the *Product Plots* framework (Wickham and Hofmann, 2011)

These charts have different strengths and weaknesses. For example, fluctuation diagrams and equal bin size plots are useful for emphasising patterns related to data sparsity, including empty combinations Hofmann et al. (2000). Sönning and Schützler (2023) suggest that double-decker plots may be preferable to traditional mosaic plots when a dataset comprises three or more variables, as this avoids comparisons of non-aligned tile lengths. In turn, *rmb plots* (Section 3.4.1.1) are generally a better option than double-decker plots when both the frequencies of combinations of explanatory variables vary considerably, and the conditional relative frequencies of response categories, or the difference between them, is small (Pilhöfer and Unwin, 2013).

Some techniques within the mosaic family represent observed frequencies less directly than traditional mosaic plots, either by emphasising expected frequencies (e.g., sieve plots) or deviations from expected independence (e.g., association plots). Association plots and fluctuation diagrams were classified within the mosaic family, rather than the size-encoding family, since both the width and heights of the bars vary.

3.4.2.3 Implicit Tree Family

Implicit tree visualisations constitute another relevant type of space-filling technique. These visualisations represent hierarchies without explicitly showing parent-child relationships, instead using positional encodings of nodes, such as node overlap or containment (Schulz et al., 2010). The techniques that work best for multivariate categorical data emphasise the *size* of nodes within a visualisation, corresponding to combination frequencies, more than they do the *structure* of the tree.

A prominent technique in the Implicit Tree family is the *sunburst diagram* (Stasko and Zhang, 2000). This technique shows the proportion of different categories and combinations of categories via a series of concentric rings. Each ring corresponds to a different variable, with the angle of each slice being proportional to the frequency of the category (first level) or combination of categories (subsequent levels) that it represents. Figure 3.9 illustrates two examples for the Titanic dataset. Outer levels are conditioned on inner levels, effectively showing conditional relative frequencies. If too many variables are

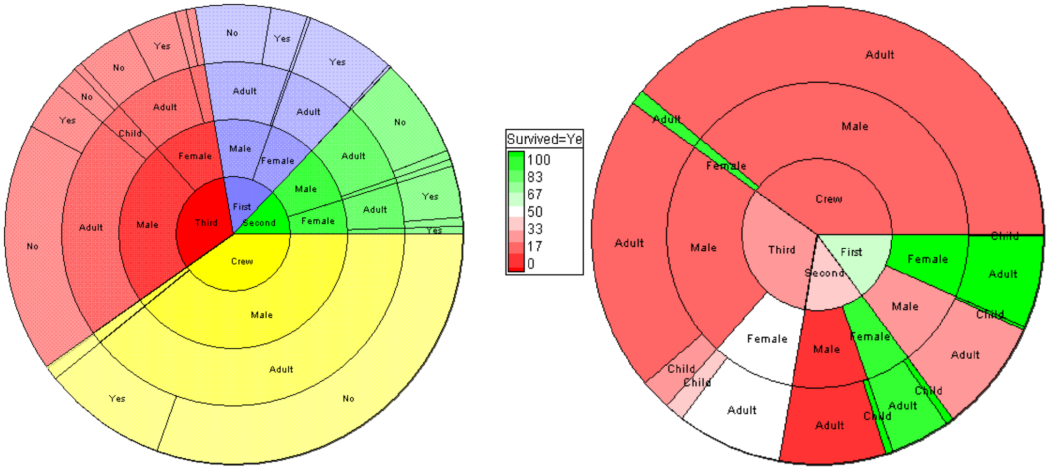


Figure 3.9: Left: Sunburst diagram showing all four variables of the Titanic dataset. Right: One of the variables (Survived) is removed from the sunburst itself and instead emphasised using colour (Clark, 2006).

shown, the slices invariably become thin and unreadable. However, zoomable versions of sunburst diagrams can help to accommodate a larger number of categories and variables.

Other implicit tree techniques that can be applied to multivariate categorical data include:

- *Categorical Treemaps* (Johnson, 1993), including *CatTrees* (Kolatch and Weinstein, 2001)[2]
- *Voronoi treemaps* (Balzer and Deussen, 2005)
- *Circular treemaps* (Wang et al., 2006), also called *circle packing*, *packed circles* and *pebbles*
- *Icicle plots* (Kruskal and Landwehr, 1983)
- *Radial Icicle Trees* (Jin et al., 2023)
- *Hi-D Maps* (Reza and Watson, 2019)

3.4.2.4 Miscellaneous Space-Filling Techniques

A handful of space-filling techniques for categorical data do not fall neatly into the above families. These include *Karnaugh-Veitch-Maps* (*KVMaps*; May et al., 2007; 2010), *Nested Rings* (*NRV*; Vivacqua and Garcia, 2008), the *Attribute Map View* (Liu et al., 2009) and *concentric pie charts* (Wickham and Hofmann, 2011). On the surface, Nested Rings appear similar to sunburst diagrams but they are not recursively subdivided; instead, they show marginal (univariate) frequencies for each variable. This is also how the Attribute Map View differs from a regular treemap.

3.4.3 Table Techniques

Techniques in the table family utilise visual encodings within each cell of a table, such as colours and bars, instead of displaying only raw text. We divide these techniques into three sub-categories: *tabular, graphical contingency tables* and *pairwise matrices*. Tabular techniques and pairwise matrices are generally well-suited to heterogeneous data, while graphical contingency tables are designed for purely categorical data.

3.4.3.1 Tabular Family

Tabular visualisations leverage the intuitiveness of a spreadsheet, with each row (or column) representing a data item or aggregate, and each column (or row) representing a variable. Prominent examples of tabular techniques

[2]Although devised independently, these are similar to mosaic plots.

that accommodate multiple categorical variables include *Table Lens* (Rao and Card, 1994) and *Taggle* (Furmanova et al., 2020). Table Lens displays each categorical variable as a 'blip'—a horizontal coloured line aligned with the category's name—while Taggle provides additional multi-form encodings, including a 'matrix' arrangement and 'colour' square with an adjacent text label. Both techniques support common spreadsheet operations, such as sorting and filtering, as well as overview and detail displays. Another notable technique in this family is the *Tableplot* (Tennekes and de Jonge, 2013; Tennekes et al., 2013). This technique requires a continuous variable for sorting but supports several high-cardinality categorical variables, as shown in Figure 3.10.

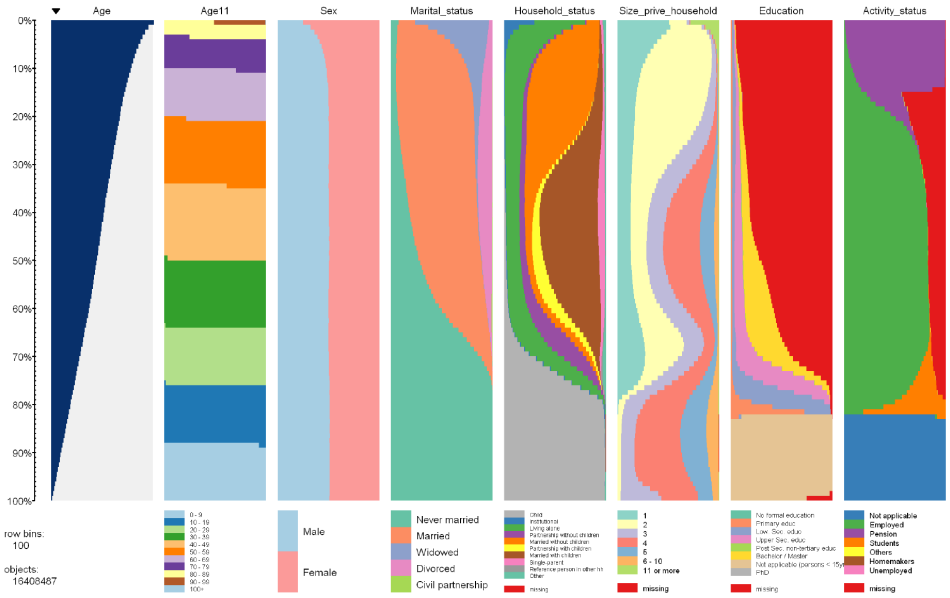


Figure 3.10: Tableplot of census data showing seven categorical variables (Tennekes and de Jonge, 2013). The data and legend are not the focus here; the figure is simply included to provide the general look and feel of this technique.

3.4.3.2 Graphical Contingency Tables

Graphical contingency tables provide a visual representation of an n-way table. The arrangement of variables and categories within the table affects the patterns that can be seen.

Notable examples of techniques in this family are *dimensional stacking* (LeBlanc et al., 1990), *colour-coded text tables* and *balloon plots* (Jain and Warnes, 2006). Dimensional stacking produces a heatmap, like the one in Figure 3.11, by embedding grids within grids. The heatmap contains one cell for each possible combination of categories, and is helpful for identifying clusters, outliers and patterns in the data. Dimensional stacking can be implemented

Figure 3.11: Dimensional stacking showing bacteria resistance against eight antibiotics, labelled a_1 to a_8. Adapted from Tominski and Schumann (2020).

in spreadsheet software using *Pivot tables* in conjunction with conditional formatting.

In terms of scalability, dimensional stacking should be limited to nine variables, each with no more than roughly five categories (Hoffman and Grinstein, 2001). Balloon plots are similar to dimensional stacking, but they display coloured circles in each cell, which are sized according to frequency. The colour of the circles can either redundantly encode this value or highlight the categories of a particular variable of interest.

3.4.3.3 Pairwise Matrices

The final type of table technique that we identified is pairwise matrices. These techniques feature a plot for each pair of variables in the data, thereby displaying all possible bivariate relationships. Univariate summaries can optionally be shown along the main diagonal. Examples that support purely categorical data are the *Heatmap Matrix* (Rocha and da Silva, 2018, 2022) and *Mosaic Matrix* (Friendly, 1999), while the *GPLOM* (Im et al., 2013) and *Generalized Pairs Plot* (Emerson et al., 2013) are suitable for mixed data. The GPLOM uses a heatmap matrix for pairs of categorical variables, whereas the GPP offers a choice between a mosaic plot, fluctuation diagram, or faceted bar chart. The GPLOM provides the richest interaction out of these techniques.

A shared property of most pairwise matrices—apart from displays involving mosaic plots—is that they are fully symmetrical. This means that only half

of the matrix needs to be displayed. Nevertheless, it can be beneficial to keep the full display, so that panels relating to each variable can be identified in a straight line, with the user focusing on either rows or columns. The Heatmap Matrix differs from the other techniques in that it allocates a fixed amount of space per category, rather than per variable. This enhances readability when a small number of variables have more than five categories. One limitation of pairwise matrices is that they do not display multivariate relationships directly, though this can be accomplished via brushing and linking across panels. In all cases, reordering rows and columns can be helpful for identifying relevant patterns.

3.4.4 Glyph Techniques

Glyphs and icons can also be used to represent categorical data, including pictorial, associative and geometric symbols (Robinson et al., 1984, p. 288). When designing glyphs for categorical data, it is important to consider the number of variables and internal categories to be represented, as well as suitable combinations of variables and encodings. Individual glyphs may be created for individual items, or for each combination of categories. In the latter case, the frequency of each combination can be mapped to the size of the glyph (e.g., Dennig et al., 2024). Incorporating a reference glyph can aid viewers in decoding the mappings (Maguire et al., 2012). Additionally, sorting glyphs by one or more variables can be beneficial (Chung et al., 2015; Ancker et al., 2011).

Examples of glyph techniques include *Star plots* (Coekin, 1969), *Autoglyphs* (Beddow, 1990) and *Chernoff faces* (Chernoff, 1973), but see Ward (2002) for a detailed list. Chernoff faces involve mapping variables to facial features, such as mouth size and face colour, and they support roughly a dozen variables. They are well suited to low-cardinality categorical data where not many values have to be discriminated. Disadvantages of Chernoff faces include that the mappings can be unnatural, and may convey unintended emotional states. De Soete and Do Corte (1985) found that only some facial features were clearly perceptible to users. Consequently, they recommended using those features for encoding the most important variables.

An advantage of glyphs over other techniques is that they enable designs that leverage metaphors and semantic relations. Domain-specific encodings promote 'natural mappings' (Siirtola, 2005), which increases understanding of glyphs and their memorisation (Maguire et al., 2012; Borgo et al., 2013). An example of metaphorical glyphs, applied in the context of hearing loss context, is shown in Figure 3.12 (Ramos et al., 2023).

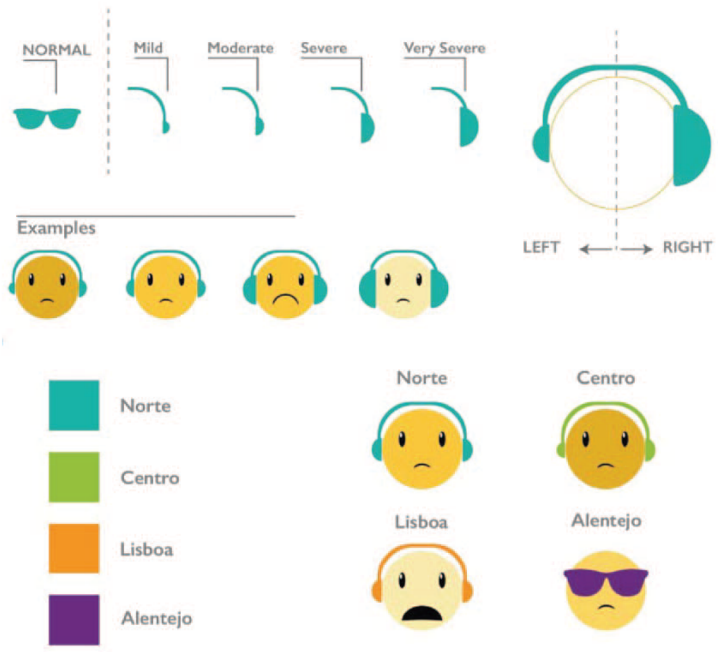


Figure 3.12: Metaphorical 'emoji' glyph, where each glyph represents a person (Ramos et al., 2023). Several categorical variables are encoded: hearing loss in left and right ears (sunglasses or headphones), region (colour), ear test appointment status (facial expression) and age (face colour). © 2023 IEEE.

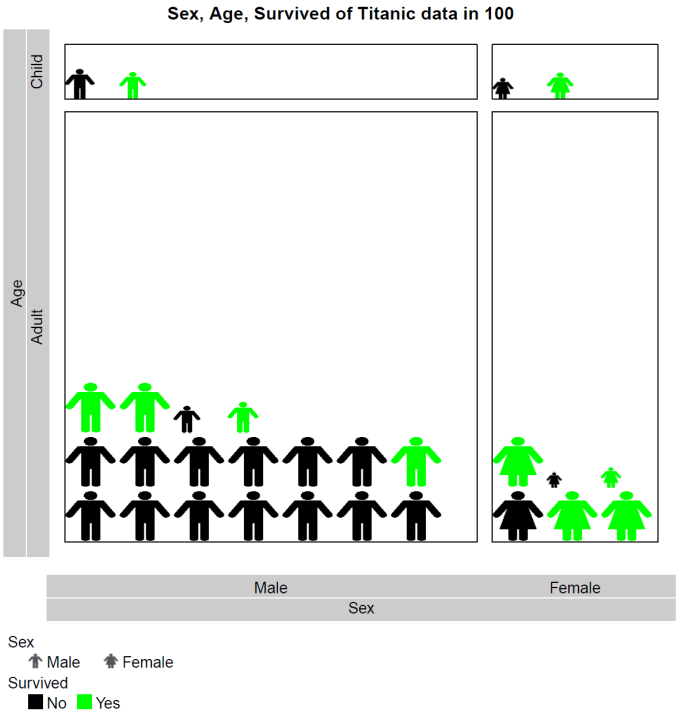


Figure 3.13: Icon plot of the Titanic dataset where each full-sized item represents 100 people (Wolf, 2021).

When representing categorical data, glyphs are typically only feasible if there is a relatively small number of categories per variable. Other, more general disadvantages of glyphs relate to their size, the limited capacity of visual channels and the cognitive demand they place on the viewer (Borgo et al., 2013).

Alternatively, instead of using complex glyphs, simple icons can be organised within a grid display, typically just varying the use of colour and/or shape. This approach is exemplified by *frequency grids* (Micallef et al., 2012), *Gatherplots* (Park et al., 2023), and *icon plots* (Wolf, 2021). Figure 3.13 illustrates an icon plot of the Titanic dataset, in which each full-sized icon represents 100 people. Such plots are relatively simple to interpret.

3.4.5 Miscellaneous Techniques

Several other frequency-based (CatViz) visualisation techniques for categorical data do not fit into the above groups. These include but are not limited to:

- *Cleveland dot plots* (Cleveland, 1984) and *lollipop charts*
- *Spreadplots* as implemented in ViSta (Valero-Mora et al., 2003)
- *Granular Representation* (Shiraishi et al., 2009)
- *Kinetica* (Rzeszotarski and Kittur, 2014)
- *Cobweb diagrams* (Upton, 2000)
- *CatNetVis* (Thane et al., 2023)
- *Conditional Inference Trees* (Hothorn et al., 2006)
- Multivariate bar charts with an explicit *tree diagram* (Kosara, 2007)
- *ContingencyWheel*(Alsallakh et al., 2011) and *ContingencyWheel++* (Alsallakh et al., 2012)
- *Worlds within worlds* (Feiner and Beshers, 1990)
- *Trilinear plots* (Allen, 2002)
- *Tetrahedrons* (Fienberg and Gilbert, 1970)
- Various *set* and *hypergraph* representations, where categories are represented as sets (e.g., *RectEuler*; Paetzold et al., 2023) or hyperedges (e.g., *PAOHVis*; Valdivia et al., 2019)

CatNetVis (Thane et al., 2023), shown in Figure 3.14, represents categories as nodes in a force-directed network. Connected nodes are attracted to each other, and non-connected nodes are repelled. An advantage of this layout is that no order needs to be specified for either the categories or variables. The size of each node represents its frequency, while its colour is determined by the mode response category. Node labels show the name of the corresponding

category and variable, with font size denoting entropy. The width of each edge is proportional to the overlap between the corresponding categories, as calculated by the Jaccard Index. Edges can be filtered by entropy to reduce clutter and home in on specific communities, aided by zooming and tooltips. With these interactive capabilities, CatNetVis can be used to explore dozens of variables and hundreds of categories.

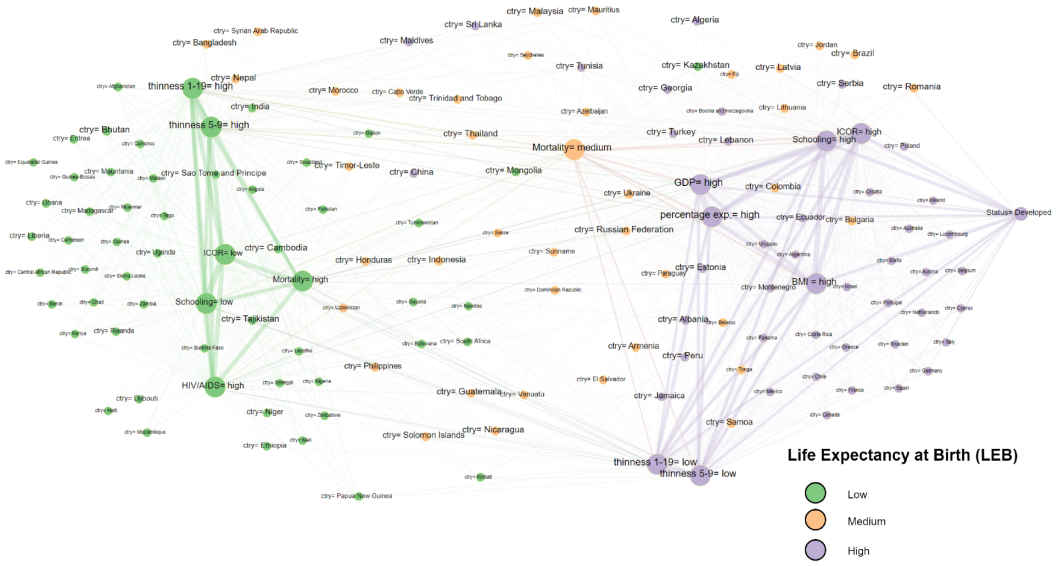


Figure 3.14: CatNetVis showing life expectancy data with a filter applied (Thane et al., 2023). There are two main communities, relating to underdeveloped countries (left) and developed countries (right).

3.4.6 Projection Techniques

Following the QuantViz approach mentioned in Section 3.3, techniques in the projection family map high-dimensional data into a low-dimensional space. The goal is to preserve relationships in the data, such as distances, similarities and associations between categories. At the heart of this approach are two key steps: a dimensionality reduction technique *transforms* categories into numbers, and a visualisation technique *represents* the result of this transformation. For a detailed review of projection techniques, see Johansson Fernstad (2011).

The most well-known dimensionality reduction techniques for categorical data are *Correspondence Analysis* (*CA*; Greenacre, 2017) for only two categorical variables, and *Multiple Correspondence Analysis* (*MCA*; Tenenhaus and Young, 1985) for greater numbers of variables. These dimensionality reduction techniques have many different names and variations.

Popular choices for visualising the results of Correspondence Analysis are *CA Maps* and *Biplots* Gabriel (1971), which are both types of scatterplots. Other visualisation techniques used for CA and MCA include:

- *Contribution Biplots* (Greenacre, 2013)
- *Moon Plots* (Bock, 2011)
- *Voronoi Diagrams* (Broeksema et al., 2013); see especially the *Categorical Data Map* (Dennig et al., 2024)
- *Andrews Curves* (Rovan, 1994)
- *Dendrograms* (Beh and Lombardo, 2014)
- *Z-Plots* (Choulakian et al., 1994)
- *Chernoff Faces* (Beh and Lombardo, 2014)

While Correspondence Analysis and Multiple Correspondence Analysis are useful for capturing structure in high-dimensional categorical datasets, they have a number of drawbacks. Both techniques are difficult for non-experts to interpret, they do not display frequency-related information, or convey the reasons *why* items belong to particular clusters. Furthermore, CA and MCA quickly become cluttered when the number of categories increases, since the individual category labels are usually shown next to the points themselves. When there are large numbers of variables in MCA, it is also difficult to determine which categories belong to which variables. These plots result in overlapping labels when there are many categories, and are generally sensitive to outliers.

Some interactive tools combine both stages of the QuantViz process in a user-controlled way. MiDAVisT (Johansson Fernstad, 2009; Johansson and Johansson, 2010), shown in Figure 3.15, is one such approach. The figure has been chosen to show all views at once, thereby highlighting the interactive capabilities; the details of the categories and text are not important. This tool provides suggestions for numeric representations to the user, which they can adjust interactively. The user can then explore the results using a range of visualisation methods in multiple coordinated views, as well as algorithmic analyses, such as clustering and correlation analysis.

Figure 3.15: Multiple view environment within MiDAVisT (Johansson and Johansson, 2010), consisting of a scatterplot matrix (top left), Table Lens visualisation (top right) and parallel coordinates display (bottom).

3.5 Analysis Tasks for Categorical Data Visualisation

Before moving on to a general comparison of the techniques in our taxonomy and a discussion of avenues for future work, we provide an overview of nine kinds of analysis tasks associated with categorical data visualisation. An awareness of such tasks is helpful for designing, improving, evaluating and comparing categorical visualisation techniques. We note that not all of these categories are applicable to all techniques.

Overview tasks constitute a useful first step in any analysis of categorical data: these include determining the total (or selected) number of data items, variables and categories in a dataset, as well as inspecting the distribution of variable cardinalities. Users should be familiar with the basic structure and characteristics of a dataset before proceeding with their analysis.

Drawing inspiration from Unwin and Pilhoefer (2020), *missing value tasks* are concerned with obtaining a summary of missing values in the data, so that these can be dealt with appropriately. Missing value tasks for categorical data may relate to variables or data items (records). For example, a user may wish to summarise the number of missing values across all variables, or determine the number of data items that are incomplete, before filtering or removing

them. Missing value tasks may be seen as a subset of overview tasks, since they involve gaining a preliminary understanding of the structure of the data.

Identification tasks focus on contextualising individual data units. Examples include determining which variable a particular category belongs to, determining which categories belong to a particular variable, and identifying any response variables within the dataset. An example involving multiple variables is identifying which categories are present in a given combination of categories, such as a particular tile in a mosaic plot.

Frequency tasks, which were a key part of Johansson Fernstad and Fernstad's (2011) study, play a crucial role in categorical data analysis. These tasks involve determining, comparing and ranking the frequency of particular categories or combinations of categories. The tasks may be univariate (involving *marginal* frequencies of one or more variables) or multivariate (involving *cross-tabulation* of two or more variables). Examples of univariate tasks are inspecting the marginal distribution of each variable, determining the n-th most (or least) frequent category across the entire dataset and ranking all categories within a particular variable by frequency. Examples of multivariate tasks are comparing the joint frequencies of two or more combinations of categories, determining conditional frequencies of a target variable for each combination of explanatory variables, and identifying the number of empty combinations involving n variables.

Similarity tasks, also explicitly mentioned by Johansson Fernstad and Fernstad (2011), involve identifying structural patterns and clusters within the data. These tasks operate at both the category and variable levels. Clustering seeks to identify groups of items that are similar to each other and different to items belonging to other clusters. Examples of similarity tasks include identifying the n most similar categories within a given variable, finding clusters of combinations of categories, and identifying the n most similar variables to a given variable. In certain contexts, it may also be helpful to identify which variable is *least* similar to all others.

To support these tasks, various similarity measures can be used, such as the overlap similarity measure, Jaccard index, Cosine distance and mutual information (Boriah et al., 2008 discuss additional measures). The most effective approach for computing some of these measures involves converting variables into multidimensional binary attributes through one-hot encoding, then comparing the resultant vectors across variables.

Co-occurrence tasks combine elements of the previous two task types, investigating conditions under which two or more categories appear together within the dataset. Examples include finding categories across a given set of variables that occur together at least $p\%$ of the time, and finding n categories from any other variables that a given category occurs with most.

Association tasks explore the relationships between variables or categories, aiming to determine if and how they are associated. Investigating category frequencies by themselves can be misleading if uncorrelated variables exist. Examples of association tasks include discerning global associations between variables, detecting individual associations between categories of different variables, and identifying one-way dependencies where one category nearly always occurs with another, but not *vice versa*. Several association measures are available for analysing categorical and ordinal data, including Cramer's V and the Goodman-Kruskal tau index (Goodman et al., 1979).

Deviation tasks involve determining the extent to which the observed data deviate from expected values. They can be helpful for identifying patterns and outliers in the data, and determining the lack-of-fit of a log-linear model. Common examples of deviations include Pearson residuals, Standardised residuals and Deviance residuals. Typical tasks are finding the combination with the smallest/largest deviation, finding the deviation of a given combination of categories and examining the distribution of deviations for all combinations involving n variables.

Finally, *data item tasks* are related to the records in a dataset, and are only applicable when a dataset contains individual identifiers (e.g., passenger names are included in some versions of the Titanic dataset). Example tasks include looking up a data item based on its identifier, comparing categories among two or more data items, and summarising category counts for a particular group of data items. Currently, only a few categorical visualisation techniques support analysis of individual data items.

3.6 Discussion and Future Work

We now consider general strengths and weaknesses of the visualisation families reviewed in this chapter, as well as possible avenues for future work in the area of categorical data visualisation. The different families of techniques in our taxonomy have different strengths and weaknesses, affecting their suitability for different contexts and analysis tasks. An overview of key points is provided in Table 3.2.

Table 3.2: Comparison of visualisation families.

Family	**Strengths**	**Weaknesses**
Size-encoding (e.g., bar charts, pie charts)	• Intuitive (no training required) • Bars support precise comparisons • Can be faceted to show extra variables • Useful for part-to-whole comparisons for a single variable	• Limited to few categories per variable or few variables • Wedges suffer from perceptual distortions • Linking becomes complicated with many variables
Space-filling (e.g., ParSets, mosaic plots)	• Area (spatial regions) well-suited to categorical data • Optimize the space used • Useful for emphasising a response variable • Relatively independent of number of data items	• Quickly become cluttered • Different orders vastly change the display / sensitive to ordering • Suffer from visual interference (e.g., line-crossings)
Glyphs (e.g., Chernoff faces, metaphoric glyphs)	• Visually emphasise items as individual objects • Can use semantically meaningful representations • Suitable for both dense and sparse structures	• Poor scalability (if using one glyph per item) • Usually requires carefully chosen variable-to-glyph mapping • Learning and memorisation can be cognitively demanding • Not all glyphs suitable for nominal variables
Table (e.g., Table Lens, Taggle)	• Utilise a familiar, spreadsheet-like layout • Fairly scalable in terms of both categories and variables • Direct representation of individual records • Well-suited for heterogeneous data • Pairwise matrices provide a compact overview	• Pairwise matrices limited to bivariate relationships • May confuse frequency in heatmap with similarity • Desired properties not always possible (many frequent combinations)

Continued on next page

Table 3.2 continued from previous page

Family	Strengths	Weaknesses
Projection (e.g., M/CA, biplots)	• Excel at similarity tasks • Can handle many variables • Useful for cluster analysis • Any visualisation technique for numeric data can be used	• Lack of frequency information • Not easily interpretable • Sensitive to outliers • Distortion (e.g., horseshoe) effects • Cluttered when there are lots of categories • First two dimensions may not capture sufficient variance

Regarding future work, the analysis tasks outlined in the previous section can be used to identify gaps in existing work, such as the lack of explicit consideration of missing values within most categorical visualisation techniques. There is also potential for visualising the results of a wider range of similarity and association measures for sets of two or more categorical variables.

A major limitation of the reviewed techniques is their general lack of scalability. Most techniques scale exponentially when a categorical variable is added, quickly leading to visual clutter and increased computation time. High-cardinality variables are also problematic, not least because channels like colour can only show about 6-8 categories effectively. On the other hand, the scalability of categorical visualisation techniques is relatively independent of the number of data items, except when these are displayed individually, as with various table and glyph techniques. Some of the reviewed techniques accommodate only a small number of variables, while others support multiple variables but only consider pairwise relationships. Rarely does a technique enable visualisation of relationships between many variables and categories simultaneously. However, techniques like CatNetVis and the Categorical Data Map are promising recent developments. Even so, there remains a need for more powerful categorical visualisation techniques that make use of visual channels in perceptually efficient ways.

Crucially, the field would benefit from more comprehensive user studies that focus specifically on multivariate categorical data. Our review of the literature suggests that there have been few developments in this area since this gap was identified by Johansson Fernstad and Johansson (2011), apart from the study carried out by Hofmann and Vendettuoli (2013). The proposed task and technique classifications in this survey paper provide a framework for designing such studies: representative techniques from different groups in our taxonomy

(Section 3.3) can be compared with respect to key analysis tasks (Section 3.5) using datasets of varying complexity. Online user studies for multivariate categorical data could be facilitated by the *ReVISit* software framework (Ding et al., 2023).[3]

Only a small proportion of the reviewed techniques have publicly available implementations that do not require programming skills and which allow users to analyse their own datasets. Few interactive tools are available for techniques that were proposed more than ten years ago (e.g., we could not find implementations for Nested Rings, Granular Representation or KVMaps). Even some more recent techniques suffer from this problem (e.g., the Heatmap Matrix and CatNetVis). User-friendly tools for other techniques—like Parallel Sets and ContingencyWheel++—were previously available but are no longer maintained, making them less accessible, or indeed inaccessible, to non-computer scientists. The development of modern, code-free tools is needed to democratise access to these techniques.

Existing visualisation techniques offer significant potential for enhancement through increased interaction. For example, allowing flexible changes to data mapping can increase the readability of glyph-based techniques. Similarly, the ability to reorder variables is crucial for techniques where a hierarchical structure is imposed, such as Parallel Sets, mosaic plots and sunburst diagrams, since different orderings can drastically alter the display. For such techniques, providing an interactive, separate view of the imposed tree structure could facilitate understanding and exploration of different configurations of the visualisation. This could be implemented as a classic tree diagram with drag-and-drop functionality.

More generally, since it does not always make sense to incorporate all variables at once (Theus, 2008), it is beneficial to allow user-controlled inclusion and exclusion of individual variables. There should be flexibility to (re-)display variables that are not currently visible, unless the user has explicitly removed them from the dataset.

Related to this, very few existing techniques for categorical data integrate multiple coordinated views. In general, combining different representations can highlight different aspects of the data, provided the display is not overly cluttered. To our surprise, we only identified two techniques that combine CatViz and QuantViz representations (ViSta spreadplots and the Categorical Data Map). Given that these two approaches are useful for different analysis tasks (Johansson Fernstad and Johansson, 2011), connecting them in visuali-

[3]See `https://revisit.dev/`

sation systems offers potential to harness their relative strengths. For example, it would be interesting to be able to view Parallel Sets and CA plots side-by-side, and to enable linked interactions between them. Even if plots cannot be shown side-by-side, due to lack of screen space, it is helpful to be able to switch between different representations while preserving selections.

Furthermore, apart from some tabular and glyph-based techniques (e.g., Table Lens, Chernoff faces), few categorical visualisation techniques support the display or analysis of individual records, as per the *data item* tasks detailed in Section 3.5. Providing access to individual identifiers and any additional string-type (text) variables in the raw data enables users to address micro-questions about specific records. While it is possible to display limited text about each data item in area-proportional visualisations of categorical data (e.g., as demonstrated by Brath, 2018, p. 155), a more scalable solution involves displaying the text within a coordinated table view (following Liu et al., 2009). For instance, clicking on different visual elements (e.g., a bar in a multivariate bar chart, a tile in a mosaic plot or a parallelogram in Parallel Sets) could highlight or isolate the corresponding records in the table view. This could be powerfully assisted by search functionality that targets the identifier. Many existing categorical visualisation techniques could be extended in this way.

We stated at the beginning of the chapter that, in our view, categorical visualisation techniques should support both nominal and ordinal data. Some QuantViz techniques, including several variants of Correspondence Analysis, take the order of categories into consideration (Beh and Lombardo, 2014). Surprisingly, however, we did not encounter any CatViz tools where ordinal variables were treated or displayed differently from nominal variables. For instance, it may be more appropriate to use greyscale for ordinal variables instead of colour, in accordance with perceptual guidelines (Mackinlay, 1986). Thus, exploring potential enhancements and customisations for ordinal variables is yet another avenue for future work.

3.7 Postscript

This chapter has reviewed existing techniques for visualising categorical data. After explaining our scope and methodology, we introduced a two-level technique taxonomy, providing a foundation for understanding and comparing different approaches. Situated within the established CatViz/QuantViz framework, this taxonomy outlines six distinct families: *size-encoding*, *space-filling,*

table, glyph, miscellaneous (all frequency-based), and *projection* (quantification-based). We discussed prominent examples from each family, ranging in complexity from simple bar charts to much more sophisticated tools like CatNetVis and MiDAVisT.

In Section 3.5, nine different kinds of analysis tasks for dealing with categorical data were proposed, from *overview tasks* to *frequency* and *association tasks*. This was followed by a summary of the general strengths and weaknesses of each family of techniques. Finally, we pinpointed areas for future research, emphasising the need for more scalable solutions, empirical user studies, code-free tools and enhanced interaction. We also advocated for better support for individual data items, as well as for handling ordinal variables alongside nominal ones. The remaining chapters in Part II (Chapters 4 & 5) present new and adapted techniques that seek to address some of these gaps, with a particular focus on improving scalability and interaction.

3.8 References

Agresti, A. (2012). *Categorical data analysis*. John Wiley & Sons, 3rd edition.

Agresti, A. (2019). *An introduction to categorical data analysis*. John Wiley & Sons, 3 edition.

Allen, T. (2002). Using and interpreting the Trilinear plot. *Chance*, 15(3):29–35.

Alsakran, J., Huang, X., Zhao, Y., Yang, J., and Fast, K. (2014). Using entropy-related measures in categorical data visualization. In *2014 IEEE Pacific Visualization Symposium*, pages 81–88. IEEE.

Alsallakh, B., Aigner, W., Miksch, S., and Gröller, M. E. (2012). Reinventing the Contingency Wheel: Scalable visual analytics of large categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2849–2858.

Alsallakh, B., Gröller, M. E., Miksch, S., and Suntinger, M. (2011). Contingency Wheel: Visual analysis of large contingency tables. In *EuroVA 2011*, pages 53–56. Eurographics.

Alsallakh, B., Micallef, L., Aigner, W., Hauser, H., Miksch, S., and Rodgers, P. (2016). The state-of-the-art of set visualization. In *Computer Graphics Forum*, volume 35, pages 234–260. Wiley Online Library.

Ancker, J. S., Weber, E. U., and Kukafka, R. (2011). Effect of arrangement of stick figures on estimates of proportion in risk graphics. *Medical Decision Making*, 31(1):143–150.

Balzer, M. and Deussen, O. (2005). Voronoi treemaps. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 49–56. IEEE.

Beck, F., Koch, S., and Weiskopf, D. (2015). Visual analysis and dissemination of scientific literature collections with survis. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):180–189.

Becker, R. A., Cleveland, W. S., and Shyu, M.-J. (1996). The visual design and control of trellis display. *Journal of Computational and Graphical Statistics*, 5(2):123–155.

Beddow, J. (1990). Shape coding of multidimensional data on a microcomputer display. In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pages 238–246. IEEE.

Beh, E. J. and Lombardo, R., editors (2014). *Correspondence Analysis*. Wiley Series in Probability and Statistics. John Wiley Sons, Ltd.

Bock, T. (2011). Improving the display of Correspondence Analysis using Moon Plots. *International Journal of Market Research*, 53(3):307–326.

Booshehrian, M., Möller, T., Peterman, R. M., and Munzner, T. (2011). Vismon: Facilitating risk assessment and decision making in fisheries management. Technical report, Tech. Rep. TR 2011-05. School of Computing Science, Simon Fraser University ....

Borgo, R., Kehrer, J., Chung, D. H. S., Maguire, E., Laramee, R. S., Hauser, H., Ward, M. O., and Chen, M. (2013). Glyph-based visualization: Foundations, design guidelines, techniques and applications. In Sbert, M. and Szirmay-Kalos, L., editors, *Eurographics 2013 - State of the Art Reports*, pages 39–63. Eurographics.

Boriah, S., Chandola, V., and Kumar, V. (2008). Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 2008 SIAM international conference on data mining*, pages 243–254. SIAM.

Brath, R. (2018). *Text in visualization: Extending the visualization design space*. PhD thesis, London South Bank University.

Brodbeck, D. and Girardin, L. (2019). High-d. `https://www.high-d.com/`. [Online; accessed 13-October-2023].

Broeksema, B., Telea, A. C., and Baudel, T. (2013). Visual analysis of multi-dimensional categorical data sets. In *Computer Graphics Forum*, volume 32, pages 158–169. Wiley Online Library.

Cantu, A., Micó-Amigo, M. E., Del Din, S., and Johansson Fernstad, S. (2023). Parallel Assemblies plot, a visualization tool to explore categorical and quantitative data: Application to digital mobility outcomes. In *2023 IEEE 16th Pacific Visualization Symposium (PacificVis)*, pages 21–30. IEEE.

Chernoff, H. (1973). The use of faces to represent points in k-dimensional space graphically. *Journal of the American Statistical Association*, 68(342):361–368.

Cherry, W. and Oldford, R. (2003). Picturing probability: The poverty of Venn diagrams, the richness of Eikosograms.

Choulakian, V., Lockhart, R. A., and Stephens, M. A. (1994). Cramér-von mises statistics for discrete distributions. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique*, pages 125–137.

Chung, D. H., Legg, P. A., Parry, M. L., Bown, R., Griffiths, I. W., Laramee, R. S., and Chen, M. (2015). Glyph sorting: Interactive visualization for multi-dimensional data. *Information Visualization*, 14(1):76–90.

Cibulková, J. and Kupková, B. (2022). Review of visualization methods for categorical data in cluster analysis. *Statistika: Statistics & Economy Journal*, 102(4):396–408.

Clark, J. (2006). Multi-level pie charts. `https://www.neoformix.com/2006/MultiLevelPieChart.html`.

Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician*, 38(4):270–280.

Coekin, J. (1969). A versatile presentation of parameters for rapid recognition of total state. In *Proceedings International Symposium on Man-Machine Systems, IEEE Conference Record*, volume 69.

Cohen, A. (1980). On the graphical display of the significant components in two-way contingency tables. *Communications in Statistics-Theory and Methods*, 9(10):1025–1041.

Davies, J. (2012). Parallel Sets. `https://www.jasondavies.com/parallel-sets/`. Accessed Jaunary 12, 2024.

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *Journal of Statistics Education*, 3(3).

De Soete, G. and Do Corte, W. (1985). On the perceptual salience of features of Chernoff faces for representing multivariate data. *Applied psychological measurement*, 9(3):275–280.

Dennig, F. L., Fischer, M. T., Blumenschein, M., Fuchs, J., Keim, D. A., and Dimara, E. (2021). Parsetgnostics: Quality metrics for Parallel Sets. In *Computer Graphics Forum*, volume 40, pages 375–386. Wiley Online Library.

Dennig, F. L., Joos, L., Paetzold, P., Blumberg, D., Deussen, O., Keim, D. A., and Fischer, M. T. (2024). Toward the Categorical Data Map. *Preprint*. `https://arxiv.org/pdf/2404.16044`.

Ding, Y., Wilburn, J., Shrestha, H., Ndlovu, A., Gadhave, K., Nobre, C., Lex, A., and Harrison, L. (2023). reVISit: Supporting Scalable Evaluation of Interactive Visualizations. In *IEEE Visualization and Visual Analytics (VIS)*, pages 31–35.

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning proceedings 1995*, pages 194–202. Elsevier.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., and Wickham, H. (2013). The generalized pairs plot. *Journal of Computational and Graphical Statistics*, 22(1):79–91.

Feiner, S. K. and Beshers, C. (1990). Worlds within worlds: Metaphors for exploring n-dimensional virtual worlds. In *Proceedings of the 3rd annual ACM SIGGRAPH symposium on User interface software and technology*, SIGGRAPH, pages 76–83.

Fienberg, S. E. (1975). Perspective canada as a social report. *Social Indicators Research*, 2:153–174.

Fienberg, S. E. and Gilbert, J. P. (1970). The geometry of a two by two contingency table. *Journal of the American Statistical Association*, 65(330):694–701.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200.

Friendly, M. (1995). A fourfold display for 2 by 2 by k tables. Technical report, Technical Report 217, Psychology Department, York University.

Friendly, M. (1998). Conceptual models for visualizing contingency table data. In *Visualization of categorical data*, pages 17–I. Elsevier.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Journal of Computational and Graphical Statistics*, 8(3):373–395.

Friendly, M. (2006). A brief history of data visualization. In Chen, C., Härdle, W., and Unwin, A., editors, *Handbook of Computational Statistics: Data Visualization*, volume III. Springer-Verlag, Heidelberg.

Friendly, M. and Meyer, D. (2015). *Discrete data analysis with R: visualization and modeling techniques for categorical and count data*, volume 120. CRC Press.

Furmanova, K., Gratzl, S., Stitz, H., Zichner, T., Jaresova, M., Lex, A., and Streit, M. (2020). Taggle: Combining overview and details in tabular data visualizations. *Information Visualization*, 19(2):114–136.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application

to principal component analysis. *Biometrika*, 58(3):453–467.

Goodman, L. A., Kruskal, W. H., Goodman, L. A., and Kruskal, W. H. (1979). *Measures of association for cross classifications*. Springer.

Greenacre, M. (2013). Contribution biplots. *Journal of Computational and Graphical Statistics*, 22(1):107–122.

Greenacre, M. (2017). *Correspondence analysis in practice*. CRC press.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In *Computer science and statistics: Proceedings of the 13th symposium on the interface*, pages 268–273. Springer.

Hoffman, P. E. and Grinstein, G. G. (2001). A survey of visualizations for high-dimensional data mining. In *Information visualization in data mining and knowledge discovery*, pages 47–82.

Hofmann, H. (2001). Generalized odds ratios for visual modeling. *Journal of Computational and Graphical Statistics*, 10(4):628–640.

Hofmann, H. (2006). *Multivariate Categorical Data — Mosaic Plots*, pages 105–124. Springer New York, New York, NY.

Hofmann, H., Siebes, A. P., and Wilhelm, A. F. (2000). Visualizing association rules with interactive mosaic plots. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 227–235.

Hofmann, H. and Vendettuoli, M. (2013). Common angle plots as perception-true visualizations of categorical associations. *IEEE transactions on visualization and computer graphics*, 19(12):2297–2305.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674.

Huh, M. Y. (2004). Line mosaic plot: Algorithm and implementation. In *COMPSTAT 2004—Proceedings in Computational Statistics: 16th Symposium Held in Prague, Czech Republic, 2004*, Other Conference, pages 277–285. Springer.

Humayoun, S. R., Bhambri, K., and AlTarawneh, R. (2018). BiD-chord: An extended chord diagram for showing relations between bi-categorical dimensional data. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, pages 1–3.

Hummel, J. (1996). Linked bar charts: Analysing categorical data graphically. *Computational Statistics*, 11(1):23–33.

Im, J.-F., McGuffin, M. J., and Leung, R. (2013). GPLOM: The Generalized Plot Matrix for visualizing multidimensional multivariate data. *IEEE*

Transactions on Visualization and Computer Graphics, 19(12):2606–2614.

Indratmo, Howorko, L., Boedianto, J. M., and Daniel, B. (2018). The efficacy of stacked bar charts in supporting single-attribute and overall-attribute comparisons. *Visual Informatics*, 2(3):155–165.

Jain, N. and Warnes, G. R. (2006). Balloon plot. *The Newsletter of the R Project Volume 6/2, May 2006*, 6:35.

Jin, Y., de Jong, T. J., Tennekes, M., and Chen, M. (2023). Radial Icicle Tree (RIT): Node separation and area constancy. *arXiv preprint arXiv:2307.10481*.

Johansson, S. and Johansson, J. (2010). Visual analysis of mixed data sets using interactive quantification. *ACM SIGKDD Explorations Newsletter*, 11(2):29–38.

Johansson Fernstad, S. (2009). Visual exploration of categorical and mixed data sets. In *Proceedings of the acm sigkdd workshop on visual analytics and knowledge discovery: Integrating automated analysis with interactive exploration*, SIGKDD, pages 21–29.

Johansson Fernstad, S. (2011). *Algorithmically guided information visualization: Explorative approaches for high dimensional, mixed and categorical data*. PhD thesis, Linköping University Electronic Press.

Johansson Fernstad, S. and Johansson, J. (2011). A task based performance evaluation of visualization approaches for categorical data analysis. In *2011 15th International Conference on Information Visualisation*, pages 80–89. IEEE.

Johnson, B. S. (1993). *Treemaps: Visualizing hierarchical and categorical data*. PhD thesis, University of Maryland, College Park.

Karduni, A., Wesslen, R., Cho, I., and Dou, W. (2020). Du Bois wrapped bar chart: Visualizing categorical data with disproportionate values. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM.

Kolatch, E. and Weinstein, B. (2001). CatTrees: Dynamic visualization of categorical data using treemaps.

Kosara, R. (2007). Autism Diagnosis Accuracy - Visualization Redesign. `https://eagereyes.org/blog/2007/autism-diagnosis-accuracy`. [Online; accessed 13-October-2023].

Kosara, R. (2008). Treemaps. `https://eagereyes.org/blog/2008/treemaps`.

Kosara, R. (2010). Turning a table into a tree: Growing parallel sets into a purposeful project. In Steele, J. and Iliinsky, N., editors, *Beautiful Visual-*

ization, pages 193–204. O'Reilly Media.

Kosara, R., Bendix, F., and Hauser, H. (2006). Parallel Sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics*, 12(4):558–568.

Kruskal, J. B. and Landwehr, J. M. (1983). Icicle plots: Better displays for hierarchical clustering. *The American Statistician*, 37(2):162–168.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S. J., and Marra, M. A. (2009). Circos: An information aesthetic for comparative genomics. *Genome research*, 19(9):1639–1645.

LeBlanc, J., Ward, M. O., and Wittels, N. (1990). Exploring n-dimensional databases. In *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pages 230–237.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Trans Vis Comput Graph*, 20(12):1983–1992.

Liu, S., Maljovec, D., Wang, B., Bremer, P.-T., and Pascucci, V. (2016). Visualizing high-dimensional data: Advances in the past decade. *IEEE transactions on visualization and computer graphics*, 23(3):1249–1268.

Liu, Z., Stasko, J., and Sullivan, T. (2009). SellTrend: Inter-attribute visual analysis of temporal transaction data. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1025–1032.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions On Graphics (Tog)*, 5(2):110–141.

Maguire, E., Rocca-Serra, P., Sansone, S.-A., Davies, J., and Chen, M. (2012). Taxonomy-based glyph design—with a case study on visualizing workflows of biological experiments. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2603–2612.

May, T. (2007). Working with patterns in large multivariate datasets-Karnaugh-Veitch-Maps revisited. In *2007 11th International Conference Information Visualization (IV'07)*, pages 277–285. IEEE.

May, T., Davey, J., and Kohlhammer, J. (2010). Combining details of the chi-square goodness-of-fit test with multivariate data visualization. In *EuroVAST@ EuroVis*, EuroVis, pages 45–50. Eurographics.

Meyer, D., Zeileis, A., and Hornik, K. (2008). Visualizing contingency tables. *Handbook of Data Visualization*, pages 589–616.

Micallef, L., Dragicevic, P., and Fekete, J.-D. (2012). Assessing the effect of visualizations on bayesian reasoning through crowdsourcing. *IEEE transac-*

tions on visualization and computer graphics, 18(12):2536–2545.

Miyamoto, A., Allacker, K., and De Troyer, F. (2022). Visual tool for sustainable buildings: A design approach with various data visualisation techniques. *Journal of Building Engineering*, 56:104741.

Nightingale, F. (1857). *Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army*. Private Publication, London.

Paetzold, P., Kehlbeck, R., Strobelt, H., Xue, Y., Storandt, S., and Deussen, O. (2023). RectEuler: Visualizing intersecting sets using rectangles. In *Computer Graphics Forum*, volume 42 of *Computer Graphics Forum*, pages 87–98. Wiley Online Library.

Park, D., Kim, S.-H., and Elmqvist, N. (2023). Gatherplot: A non-overlapping scatterplot. *arXiv preprint arXiv:2301.10843*.

Pilhöfer, A. and Unwin, A. (2013). New approaches in visualization of categorical data: R package extracat. *Journal of Statistical Software*, 53:1–25.

Playfair, W. (1786). *The commercial and political atlas*. Wallis.

Playfair, W. (1801). *The statistical breviary*. Wallis.

Ramos, B. N., Maçãs, C., Lourenço, N., and Polisciuc, E. (2023). Towards contextual glyph design: Visualizing hearing screenings. In *2023 27th International Conference Information Visualisation (IV)*, pages 96–102. IEEE.

Rao, R. and Card, S. K. (1994). The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322.

Reza, R. M. and Watson, B. A. (2019). Hi-D maps: An interactive visualization technique for multi-dimensional categorical data. In *2019 IEEE visualization conference (VIS)*, pages 216–220. IEEE.

Riedwyl, H. and Schüpbach, M. (1994). Parquet diagram to plot contingency tables. In *Advances in Statistical Software, F. Faulbaum (Ed.).*, pages 293–299. Gustav Fischer.

Robinson, A. H., Sale, R. D., Morrison, J. L., and Muehrcke, P. C. (1984). *Elements of Cartography*. Wiley, New York.

Rocha, M. M. N. and da Silva, C. G. (2018). Heatmap matrix: a multidimensional data visualization technique. In *Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI)*.

Rocha, M. M. N. and da Silva, C. G. (2022). Heatmap matrix: Using reordering, discretization and filtering resources to assist multidimensional data analysis. In *IADIS International Conference Computer Graphics, Visualization, Computer Vision and Image Processing 2022 (part of MCCSIS*

2022), pages 11–18. MCCSIS.

Rovan, J. (1994). Visualizing solutions in more than two dimensions. *Correspondence analysis in the social sciences*, pages 210–229.

Rzeszotarski, J. M. and Kittur, A. (2014). Kinetica: Naturalistic multi-touch data visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, SIGCHI, pages 897–906.

Sanderson, D. and Peacock, D. (2020). Making rose diagrams fit-for-purpose. *Earth-Science Reviews*, 201:103055.

Schlimmer, J. (1987). Mushroom. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5959T.

Schmidt, M. (2006). Der einsatz von sankey-diagrammen im stoffstrommanagement. Technical report, Beiträge der Hochschule Pforzheim.

Schonlau, M. (2003). Visualizing categorical data arising in the health sciences using hammock plots. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*. ASA.

Schonlau, M. (2024). Hammock plots: Visualizing categorical and numerical variables. *Journal of Computational and Graphical Statistics*, 0(0):1–16.

Schulz, H.-J., Hadlak, S., and Schumann, H. (2010). The design space of implicit hierarchy visualization: A survey. *IEEE transactions on visualization and computer graphics*, 17(4):393–411.

Shiraishi, K., Misue, K., and Tanaka, J. (2009). A tool for analyzing categorical data visually with granular representation. In *Human Interface and the Management of Information. Information and Interaction: Symposium on Human Interface 2009, Held as part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009, Proceedings, Part II*, Other Conference, pages 342–351. Springer.

Siirtola, H. (2005). The effect of data-relatedness in interactive glyphs. In *Ninth International Conference on Information Visualisation (IV'05)*, pages 869–876. IEEE.

Siirtola, H. (2014). Bars, pies, doughnuts & tables–Visualization of proportions. In *Proceedings of the 28th International BCS Human Computer Interaction Conference (HCI 2014) 28*, pages 240–245. BCS.

Skau, D. and Kosara, R. (2016). Arcs, angles, or areas: Individual data encodings in pie and donut charts. In *Computer Graphics Forum*, volume 35, pages 121–130. Wiley Online Library.

Stasko, J. and Zhang, E. (2000). Focus + context display and navigation techniques for enhancing radial, space-filling hierarchy visualizations. In *IEEE Symposium on Information Visualization 2000. INFOVIS 2000. Proceedings*,

pages 57–65. IEEE.

Streit, M. and Gehlenborg, N. (2014). Bar charts and box plots: Creating a simple yet effective plot requires an understanding of data and task. *Nature Methods*, 11(2):117.

Symanzik, J., Friendly, M., and Onder, O. (2019). The unsinkable Titanic data.

Sönning, L. and Schützler, O. (2023). Data visualization in corpus linguistics: Critical reflections and future directions. In Sönning, L. and Schützler, O., editors, *Data Visualization in Corpus Linguistics: Critical Reflections and Future Directions*, number 22 in Studies in Variation, Contacts and Change in English. VARIENG, Helsinki.

Tan, P.-N., Steinbach, M., and Kumar, V. (2006). *Introduction to Data Mining*. Addison Wesley.

Tenenhaus, M. and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119.

Tennekes, M. and de Jonge, E. (2013). On the exploration of high cardinality categorical data.

Tennekes, M., de Jonge, E., Daas, P. J., et al. (2013). Visualizing and inspecting large datasets with tableplots. *Journal of Data Science*, 11(1):43–58.

Thane, M., Blum, K. M., and Lehmann, D. J. (2023). CatNetVis: Semantic Visual Exploration of Categorical High-Dimensional Data with Force-Directed Graph Layouts. In Hoellt, T., Aigner, W., and Wang, B., editors, *EuroVis 2023 - Short Papers*. The Eurographics Association.

Theus, M. (2002). Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7:1–9.

Theus, M. (2008). High-dimensional data visualization. In Chen, C.-h., Härdle, W. K., and Unwin, A., editors, *Handbook of Data Visualization*, pages 151–178. Springer, Berlin.

Tominski, C. and Schumann, H. (2020). *Interactive visual data analysis*. AK Peters/CRC Press.

Tufte, E. R. and Graves-Morris, P. R. (1983). *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT.

Unwin, A. and Pilhoefer, A. (2020). Visna—visualising multivariate missing values. In *ICML Workshop on the Art of Learning with Missing Values (Artemiss)*.

Upton, G. J. (2000). Cobweb diagrams for multiway contingency tables. *Jour-*

nal of the Royal Statistical Society: Series D (The Statistician), 49(1):79–85.

Valero-Mora, P. M., Young, F. W., and Friendly, M. (2003). Visualizing categorical data in vista. *Computational Statistics Data Analysis*, 43(4):495–508. Data Visualization.

VanderPlas, S., Ge, Y., Unwin, A., and Hofmann, H. (2023). Penguins Go Parallel: A grammar of graphics framework for generalized parallel coordinate plots. *Journal of Computational and Graphical Statistics*, pages 1–16.

Vivacqua, A. S. and Garcia, A. C. B. (2008). NRV: Using nested rings to interact with categorical data. In *Proceedings of the IADIS International Conference on Interfaces and Human Computer Interaction*, IADIS, pages 85–92.

Wang, S., Mondal, D., Sadri, S., Roy, C. K., Famiglietti, J. S., and Schneider, K. A. (2022). Set-stat-map: Extending parallel sets for visualizing mixed data. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, pages 151–160. IEEE.

Wang, W., Wang, H., Dai, G., and Wang, H. (2006). Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 517–520.

Ward, M. O. (2002). A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3-4):194–210.

Wickham, H. and Hofmann, H. (2011). Product plots. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2223–2230.

Wilkinson, L. (2006). Revising the Pareto chart. *The American Statistician*, 60(4):332–334.

Wolf, H. P. (2021). iconplot: Icon plots for visualization of contingency tables. `https://rdrr.io/cran/aplpack/man/iconplot.html`.

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, 16(3):507–525.

Zhang, C., Chen, Y., Yang, J., and Yin, Z. (2019). An association rule based approach to reducing visual clutter in parallel sets. *Visual Informatics*, 3(1):48–57.

Chapter 4

Extending the Heatmap Matrix: Pairwise Analysis of Multivariate Categorical Data

In this chapter, we propose several extensions and modifications to the *Heatmap Matrix*, which was briefly introduced in Section 3.4.3.3, underneath the *table* family. This visualisation technique is useful for exploring pairwise relationships among a larger set of categorical variables. Since the Heatmap Matrix depicts only pairwise relationships, we consider it to be a *(multiple) bivariate* technique, rather than a multivariate one. We provide a detailed description of our empirical prototype, the *Heatmap Matrix Explorer*, which helps to address the lack of relevant interactive tools identified in Chapter 3.

Publication Details

The following paper has been reproduced with minor changes to the formatting, as discussed in Section 1.4:

Trye, D., Apperley, M, & Bainbridge, D. (2023). Extending the Heatmap Matrix: Pairwise analysis of multivariate categorical data. In *2023 27th International Conference Information Visualisation (IV).* (pp. 29-36). Tampere, Finland: IEEE. https://doi.org/10.1109/IV60283.2023.00016

Abstract

Analysts are often interested in understanding the association between variables within a dataset. This paper describes a set of techniques for augmenting the Heatmap Matrix, which represents pairwise intersections of categorical variables. The proposed extensions include adapting the design and layout of the matrix to enhance its readability, expanding the number of metrics that can be presented, displaying matching records in a coordinated table view, and embedding the Chi-square test of independence. These features are demonstrated on two datasets using the empirical prototype that has been developed.

4.1 Introduction

Categorical variables are widely used in real-world datasets across a multitude of domains, ranging from business to biomedical science (Agresti, 2013). However, visualisation techniques for categorical variables have received limited attention compared to those for continuous data (Friendly, 1992), with relatively few techniques supporting the exploration of more than a handful of categorical variables at the same time (Reza and Watson, 2019). Consequently, there are clear opportunities for advancing the state-of-the-art in categorical data visualisation, including developing novel techniques and improving upon existing ones. This paper adopts the latter approach, contributing a set of extensions for enhancing the readability, functionality and scalability of the *Heatmap Matrix* (Rocha and da Silva, 2018, 2022), which represents multiple categorical variables by breaking them down into pairwise relations. The proposed extensions collectively provide more nuanced insights into the association between categorical variables, enabling the viewer to detect patterns at both a local and global level that might otherwise be missed.

Given the focus of this paper, a more detailed description of the heatmap matrix (Rocha and da Silva, 2018) is in order. This technique provides a concise visual summary of all possible two-way contingency tables for a given set of categorical variables. The plot is a matrix of heatmaps whose rows and columns are categories grouped by variable, such that each heatmap 'panel' relates to a distinct pair of variables, and each 'cell' represents the intersection of two categories. To aid readability, the categories are ordered consistently along both axes. In previous work, the heatmap matrix has only been used to show the frequency of occurrence of the corresponding categories; however, as this paper will show, other information can also be fruitfully encoded. The technique facilitates quick identification of salient patterns and values,

accentuating outliers, as well as large numbers of cells with very low or high frequencies. Since each heatmap can be taken as an independent unit, patterns can be discovered at both a local (panel) and global (matrix) level. For instance, a user may wish to analyse each panel one at a time, locate the highest and lowest values across the entire matrix, or focus on specific categories or variables of interest by isolating particular rows or columns of the matrix.

While the original heatmap matrix was static, the authors describe several interactive enhancements in later work (Rocha and da Silva, 2022). These include: four methods for reordering categories according to different seriation algorithms; filtering based on both Spearman's correlation coefficient and association rules; bucketing of continuous variables by producing bins of equal width or frequency (Dougherty et al., 1995); and the choice of a local or global colour mapping to highlight patterns within or across the matrix, respectively. However, without a publicly available prototype or a more detailed description of the user interface, it is not clear how these features are operationalised. This paper focuses predominantly on novel features that are intended to supplement, rather than replace, those mentioned by Rocha and da Silva (2022).

In terms of scalability, the size of a heatmap matrix is proportional to the number of categories in the display, with higher-cardinality variables occupying more space. The number of data items (records) has no bearing on the dimensions of the visualisation, as this is conveyed through the colour of the heatmap. Although it is theoretically possible to generate a heatmap matrix for any number of categories or variables, the visualisation is in practice restricted by the screen resolution. When drawing heatmap matrices, datasets comprising multiple high-cardinality variables pose a significant challenge (Rocha and da Silva, 2022), which is a key consideration for this work.

4.2 Related Work

Most techniques for visualising categorical data—including the heatmap matrix—are based on contingency tables (Fernstad and Johansson, 2011). Alsallakh et al. (2012) classify these methods into three main types: frequency representations, deviation representations, and intermediate representations. Frequency representations display the observed frequencies in a contingency table directly, typically using an area-proportional encoding. Prominent examples include Mosaic Plots (without residual-based shading) (Hartigan and Kleiner, 1984) and Parallel Sets (Kosara et al., 2006), which are described in further detail below. The heatmap matrix also falls under this category, though it uses

colour rather than area to encode frequency, sacrificing precision for greater scalability (Rocha and da Silva, 2022).

Deviation representations visualise differences between observed and expected frequencies. Examples include Association Plots (Meyer et al., 2003), Sieve Diagrams (Friendly, 1992) and the dot-based Contingency Wheel (Alsallakh et al., 2011). Section 4.5 shows how the heatmap matrix can be extended to support deviations as well as frequencies; the two approaches are complementary, not mutually exclusive.

Finally, intermediate representations convert categories into numerical values before visualising them. Correspondence Analysis (CA) identifies associations between the cells in a contingency table by projecting points into a low-dimensional space (Greenacre, 2017).

Matrix-based visualisations for representing every pair of variables in a dataset constitute another class of relevant techniques. These include the Scatterplot Matrix (SPLOM) for continuous data (Carr et al., 1987), together with its enhancements (Wilkinson et al., 2005; Elmqvist et al., 2008; Waskom, 2021); the Mosaic Matrix for categorical data (Friendly, 1999); and the Generalised Pairs Plot (Emerson et al., 2013) and GPLOM (Im et al., 2013) for dealing with mixed data types (first suggested by Friendly, 1999). These latter representations use different kinds of charts depending on the variable types that are present in each pair. While there is a range of options for representing two categorical variables – including Mosaic Plots (Hartigan and Kleiner, 1984), Fluctuation Diagrams (Hofmann, 2000), and Faceted Bar Charts (Becker et al., 1996), as posited by (Emerson et al., 2013) – GPLOMs use a heatmap for simplicity. However, since GPLOMs use a fixed panel size for all pairings, regardless of variable cardinality, these heatmaps are not always readable.

Mosaic Matrices (Friendly, 1999) are specifically designed for visualising pairs of categorical variables. In this representation, variables are crossed among themselves in a matrix, and a Mosaic Plot is shown in each of the resulting panels, with variable names displayed along the main diagonal. The size of each tile in a Mosaic Plot is proportional to the cell frequencies, and the tiles are often also coloured according to Pearson residuals (Friendly, 1994), yielding a blended frequency/deviation representation. These residuals provide an indication of the goodness-of-fit of the model of independence, and show which values occur more or less often than expected. Similar to GPLOMS, however, Mosaic Plots become difficult to read when visualising categorical variables with high cardinality (Im et al., 2013). Furthermore, Mosaic Matrices are generally restricted to visualising three or four variables at a time, due to

space limitations (Friendly, 1999).

Although originally intended for visualising hierarchies of variables, Parallel Sets (Kosara et al., 2006) can also be used to show pairwise relations between categorical variables (Hofmann and Vendettuoli, 2013). However, as the number of variables increases, so too does the number of repeated bands or small multiples needed to explicitly capture all possible relationships.

Reflecting on the various strengths and weaknesses of these techniques, the heatmap matrix offers a more compact alternative for exploring pairwise associations, on which this paper seeks to build.

4.3 Empirical Prototype

The following sections describe the fundamentals of the prototype that has been developed to extend the capabilities of the heatmap matrix. An overview of the *Heatmap Matrix Explorer* is given in Figure 4.1, showing the familiar Titanic dataset (Dawson, 1995). The design consists of four components: the *Matrix View* (centre) containing the heatmap; the *Selection Menu* (left-hand side) for filtering the data and merging categories; the *Main Menu* (right-hand side) for customising the heatmap; and the *Linked Table View* (bottom) showing underlying data for selected cells. While the Titanic dataset is used as the primary example throughout the paper, a second, more complex dataset is examined in Section 4.8 to provide a clearer indication of the technique's scalability.

4.4 Matrix View

At the heart of the prototype is the *Matrix View* where the main visualisation is displayed. While this view is similar to the original heatmap matrix (Rocha and da Silva, 2018), there are several key differences. In previous work, the design had a black background and variables were separated with grey grid lines; in contrast, the new design uses a white background and replaces these grid lines with white space. This helps to achieve a minimalist aesthetic that is easier on the eye (Franconeri et al., 2021). The variable groupings can be perceived solely through the spacing between panels, in accordance with the Gestalt Law of Proximity (Koffka, 1935). In addition, all cells have been given a thin white border to help distinguish individual values. Category labels for columns are rotated 45 degrees for readability, and a legend has been added to indicate the exact range of values present in the heatmap (for global

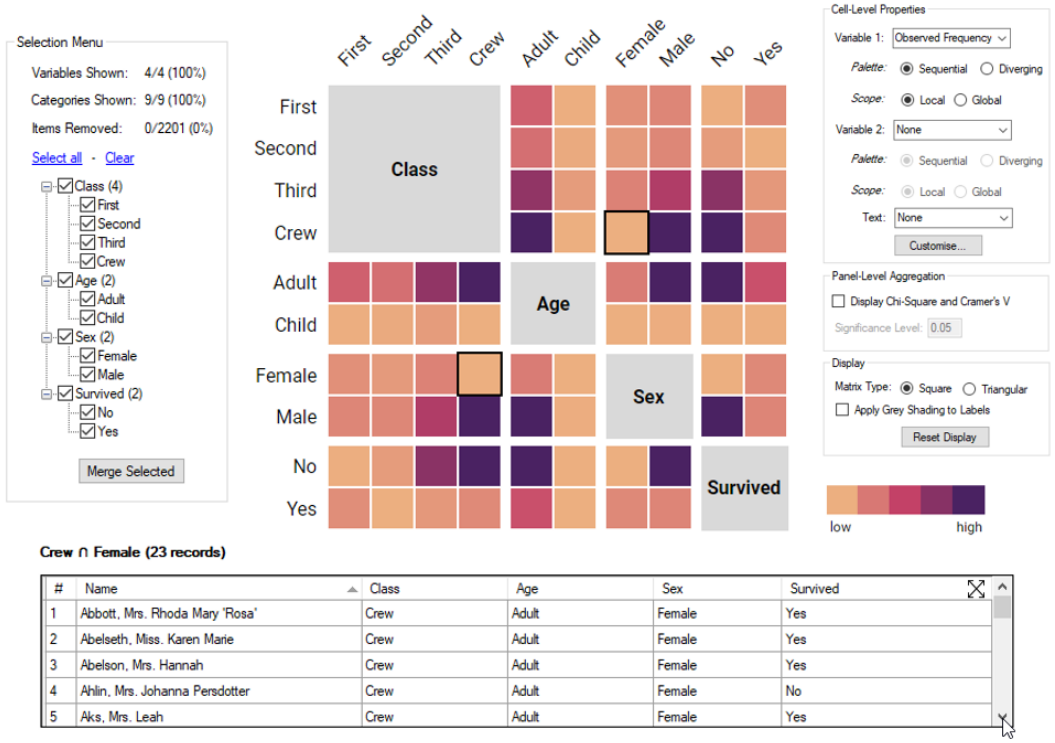


Figure 4.1: Design overview of the *Heatmap Matrix Explorer*, which represents the intersection of every pair of categories in a dataset. This example shows the Titanic dataset, consisting of 4 categorical variables and 10 categories.

mappings) or the general direction of the encoding (for local mappings). Like in the original design, all cells are square-shaped, so as not to privilege one axis (variable) over the other.

Another point of difference is the main diagonal of the matrix. In the updated design, intra-variable cell frequencies are replaced with a single grey 'box' showing the name of the corresponding variable, akin to how variables are labelled in the Mosaic Matrix (Friendly, 1999). The motivation for this is two-fold. First, it removes redundant and potentially distracting information. At least half of the cells in diagonal panels represent intersections that are structurally impossible, assuming the categories within each variable are mutually exclusive. If this is the case, the only cells that can occur represent categories' marginal frequencies; however, such data is univariate rather than bivariate, and thus has a different interpretation from the rest of the matrix. Of course, information regarding individual category frequencies may still be useful, but this can be displayed in a less obtrusive manner, by means of a tooltip; see Figure 4.3. The second reason for this change is that labelling variables along the main diagonal frees up space, since the variable names are

included within the matrix itself and do not need to be added as external row or column labels.

Two additional features supported by the prototype are tooltips and associative highlighting,[1] which work together to provide "details-on-demand" (Shneiderman, 1996). There are two different kinds of tooltips, depending on whether the user hovers over a cell (Figure 4.2) or one of the variable boxes along the main diagonal (Figure 4.3). In the former case, the tooltip displays rounded values for all seven supported metrics (Section 4.5.1), including observed frequency. Bold text is used to indicate the metric(s) that are currently encoded in the heatmap. If the user hovers over one of the variable boxes, the tooltip instead shows the distribution of category frequencies in the form of a bar chart. Categories are sorted in descending order of frequency, regardless of their position in the heatmap.

Associative highlighting helps the user to see which categories have been selected, and which variables they relate to. When the user hovers over a cell, not only does a tooltip appear, but the cell is given an orange outline, and the two related variables are highlighted orange. The corresponding row and column labels are emphasised in bold, enabling the user to accurately pinpoint their position within the matrix, which is not trivial for more complex datasets.

4.5 Main Menu

The *Main Menu* allows the user to customise the heatmap matrix in simple yet powerful ways. There are three sub-menus that control different aspects of the visualisation: cell-level properties, panel-level aggregation and general appearance of the display. The first two sub-menus cannot be used at the same time (they provide different modes for exploring the matrix), whereas the third sub-menu is compatible with both of the others, and thus always available. All features supported by these menus are novel, except for the scope setting, which was proposed by Rocha and da Silva (2022). Together, these controls provide a diverse range of complementary views that encourage users to examine the data from fresh and varied perspectives.

4.5.1 Cell-Level Properties

In the same way that contingency tables can display different measures of association, the heatmap matrix is not confined to visualising only observed frequency. The cell-level menu supports five additional metrics, from row

[1] This term is used in a different sense from Im et al. (2013).

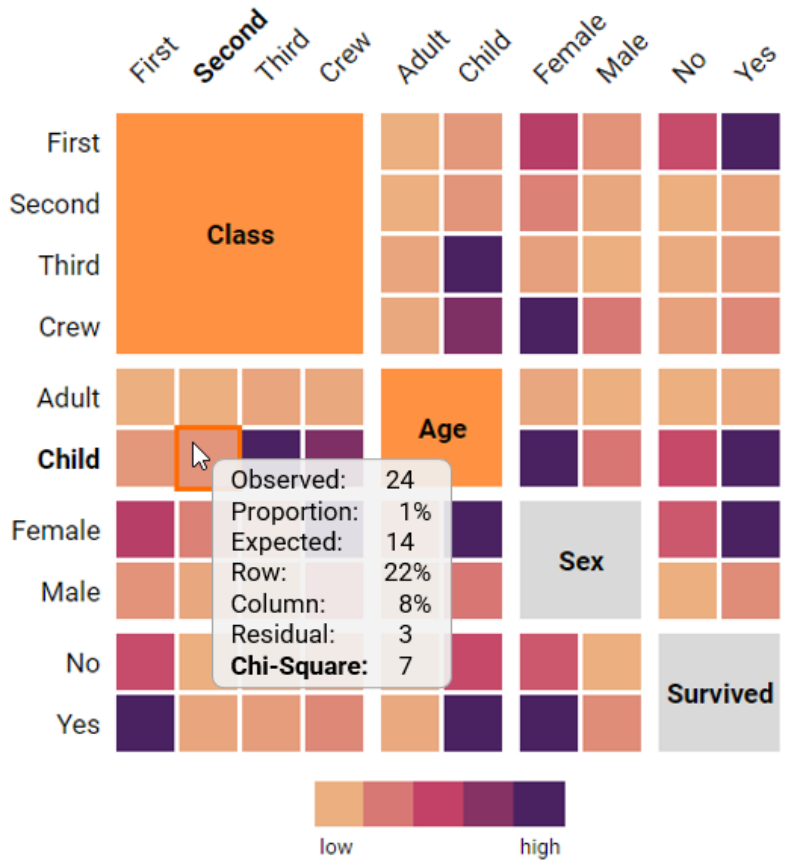


Figure 4.2: Example of a cell-level tooltip and associative highlighting. The user is currently hovering over the cell representing children in second class. The heatmap itself shows cell Chi-square values (sequential palette, local scope) for the Titanic dataset.

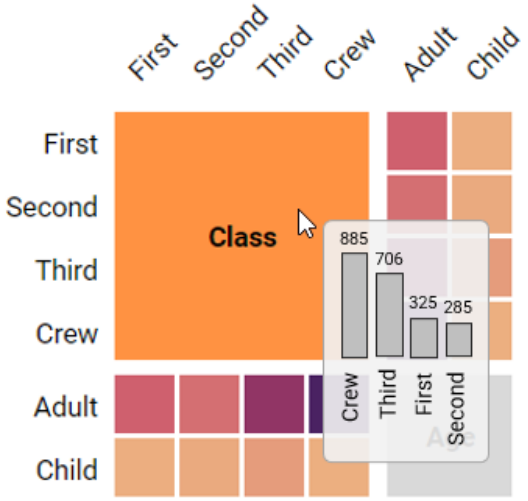


Figure 4.3: Tooltip showing a bar chart of category frequencies for the "Class" variable.

percentages to Pearson residuals, which can be used in combination to provide additional insights and supporting evidence about the nature of association between categorical variables. Within the *Heatmap Matrix Explorer*, users must select either one or two metrics to control the colour of the heatmap, and can optionally specify a third metric (or one of the same metrics) to annotate the cells with the corresponding numerical values.

For metrics defining the colour of the heatmap, the user must specify a colour palette (either sequential or diverging) and a scope (either local or global). Sequential colour palettes accentuate high values (or low values if the scale is reversed), whereas diverging colour palettes emphasise values at both ends of the spectrum. When the user chooses a local scope, the colour of each panel is scaled according to its local minimum and maximum values, rather than the extremities across the entire matrix. In general, a local mapping seems more appropriate than a global one, since, unless all variables happen to have the same cardinality, the panels in the heatmap will contain different numbers of cells, and individual cells within smaller panels are more likely to draw higher counts.

The metrics that appear in each drop-down menu are detailed below.

- *Observed Frequency* shows the frequency of occurrence in each cell, which is the same information encoded in the original heatmap matrix (Rocha and da Silva, 2018). If all categories are shown, the cells in each panel sum to the total number of data items, and each row or column sums to the category frequency. This is the default setting, useful for obtaining a preliminary overview of the data but limited in terms of measuring associations.
- *Expected Frequency* displays the quantities that would be expected in each panel if there were no association between the two variables. This is calculated by multiplying each cell's row total by its column total, then dividing by the total number of observations.
- *Row Percentages* and *Column Percentages* display the relative contribution of the observed frequency of each cell to the *local* row or column total, respectively. These metrics reveal how the categories belonging to one variable are distributed with respect to the other. For *Row Percentages*, each cell shows $P(X \mid Y)$, where X is the category on the x-axis and Y is the category on the y-axis. The correct interpretation is *what percentage of* Y *is* X? *Column Percentages* shows the inverse, i.e., $P(Y \mid X)$. The matrix generated by either metric is the transpose of the other.

- *Pearson Residuals* measure, for each cell, the magnitude and direction of the deviation from independence, adjusted for the expected variability. They are calculated by dividing the difference between the observed and expected frequencies by the square root of the expected frequency. This provides a quick visual summary of over- and under-represented pairs of categories. Cells with large residuals (in either direction) may be indicative of patterns or relationships between two variables that warrant further investigation. Pearson Residuals are best suited to a diverging colour palette (preferably one that is continuous; Zeileis et al., 2007), since this emphasises both positive and negative residuals.
- *Cell Chi-Square Values* (shown in Figure 4.2) denote the individual contribution of each cell to the overall Chi-square (χ^2) test statistic (see Section 4.5.3). The cell values are calculated by taking the squared difference between the observed and expected frequencies, and dividing by the expected frequency. A cell Chi-square value less than one means that the observed and expected frequencies are reasonably close to each other, whereas values much larger than one indicate a disparity between the two. While Pearson residuals show similar information, examining the individual cell values can help to identify outliers or unusual patterns that may not be apparent from the residuals alone, and vice versa.

One salient design consideration for any heatmap is the colour palette, which affects the range of values that can be seen (Gehlenborg and Wong, 2012; Munzner, 2014; Franconeri et al., 2021). The *Heatmap Matrix Explorer* uses sensible default colours, including Seaborn's (Waskom, 2021) perceptually linear "flare" colourmap for a single sequential metric, a blue-white-red palette for a single diverging metric, and Cynthia Brewer's nine-class bivariate maps[2] when two metrics are selected. Since the bivariate heatmap only has nine distinct values, it sacrifices precision for general readability. Nevertheless, exact values are still accessible via interactive tooltips. An example is shown in Figure 4.4, which simultaneously encodes observed frequency and Pearson residuals, such that darker cells represent more frequent values, and blue, grey and orange are used for negative, neutral and positive residuals, respectively. This encodes similar information to a Mosaic Matrix, minus the alignment of tiles, in a more scalable form.

Unlike its predecessor (Rocha and da Silva, 2018), by default, the *Heatmap Matrix Explorer* does not display numerical values in each cell. Removing the text labels makes it easier to glean general patterns, while still allowing users to

[2]http://www.personal.psu.edu/cab38/ColorSch/Schemes.html

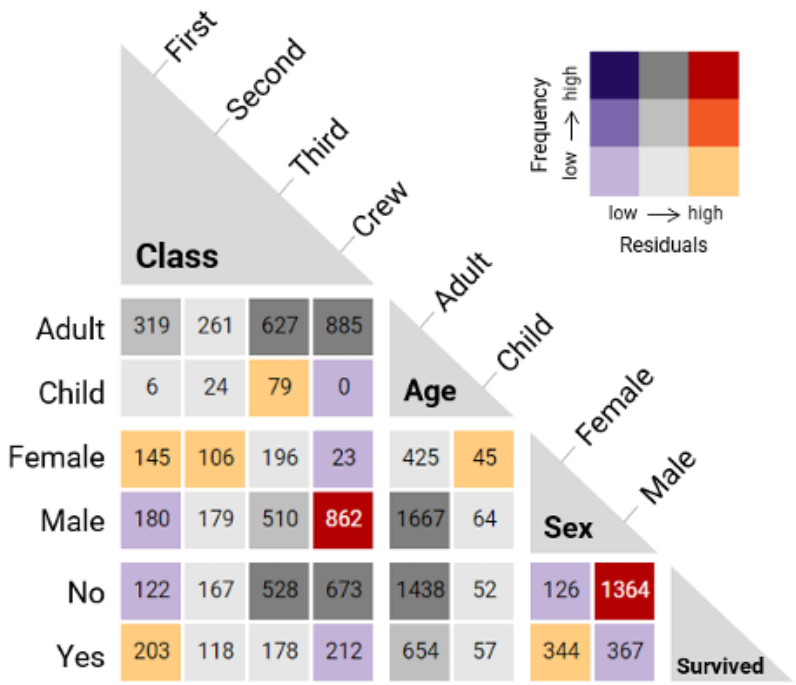


Figure 4.4: Triangular heatmap matrix showing both observed frequency (sequential palette, local scope) and Pearson residuals (diverging palette, global scope) for the Titanic dataset. The text labels also show observed frequency.

access tooltips. However, text labels may still be useful in some scenarios, such as in static (e.g., print) environments. The text drop-down menu supports the same metrics outlined above, except observed frequency is split into "Counts" and "Proportions", for which the colours are the same. Proportions display the joint probability of the corresponding categories, $P(X \cap Y)$, by dividing the counts by the total number of observations.

4.5.2 Display Settings

Among the general display settings is a "Reset" button that restores the default settings and reverts to the original dataset. There are also inputs for changing the type of matrix and the appearance of the row and column labels, which are explained below.

The prototype allows the user to switch between a square or triangular matrix (see Figure 4.2 and Figure 4.4 for examples of each). In a square matrix, all pairs of categories are represented twice, once on each axis, whereas a triangular matrix removes this redundancy. As with SPLOMs, a square matrix can be helpful for identifying patterns and trends related to particular variables of interest. This is because the user can focus on a single horizontal or vertical band of the matrix, rather than having to divide their attention

between a mixture of rows and columns, while simultaneously transposing parts of the display.

A triangular matrix, on the other hand, makes the display less cluttered and safeguards against novice users misreading the visualisation, e.g., by thinking that each value occurs twice. The outer variables in a triangular matrix are special cases as all panels appear in a straight column or row, like they would in a square matrix. As a result, the left-most variable only needs column labels and the right-most variable only needs row labels.

Regarding the appearance of labels more generally, a tickbox enables the heatmap to be drawn with or without shaded row and column labels. If applied, alternating shades of grey are used to distinguish categories belonging to adjacent variables, as shown in Figure 4.6.

4.5.3 Panel-Level Aggregation

Two widely used statistical techniques for analysing categorical data are the Chi-square (χ^2) test of independence and Cramér's V (φ_c). These techniques are often paired together to establish 1) whether there is a significant association between two variables and 2) how strong it is. Figure 4.5 shows how this information can be integrated into the heatmap itself, providing a concise visual summary of the extent to which different variables are associated.[3] This feature is inspired by correlation matrices for quantitative data (Friendly, 2002), especially those which visualise statistical significance (Wei et al., 2017). Note that, unlike other views, the cells within each panel are merged, because both measures apply to the variable level rather than the category level.

Importantly, test results are only calculated and displayed if the four basic Chi-square test conditions are thought to be satisfied. The first requirement is that both variables must be categorical. Since only categorical variables are featured in the heatmap matrix (with the possibility of binning continuous variables before the visualisation is generated), this requirement is automatically fulfilled. Secondly, observations must be independent. When the user first ticks the box to "Display Chi-Square and Cramér's V", a dialog box appears asking them to verify whether this is the case. This is the only check that cannot be automated, since it is highly context-dependent. If the user selects "No" (i.e., observations are not independent), a further message appears informing them that, unfortunately, the test cannot be applied. Otherwise,

[3]**Erratum:** The numbers given in the top left panel of Figure 4.5 (Class vs. Age) are incorrect. They should be 118.41 (3), $p < 0.001$ and $\varphi_c = 0.232$, with the panel then being the same colour as the two directly beneath it, indicating a 'Moderate' relationship.

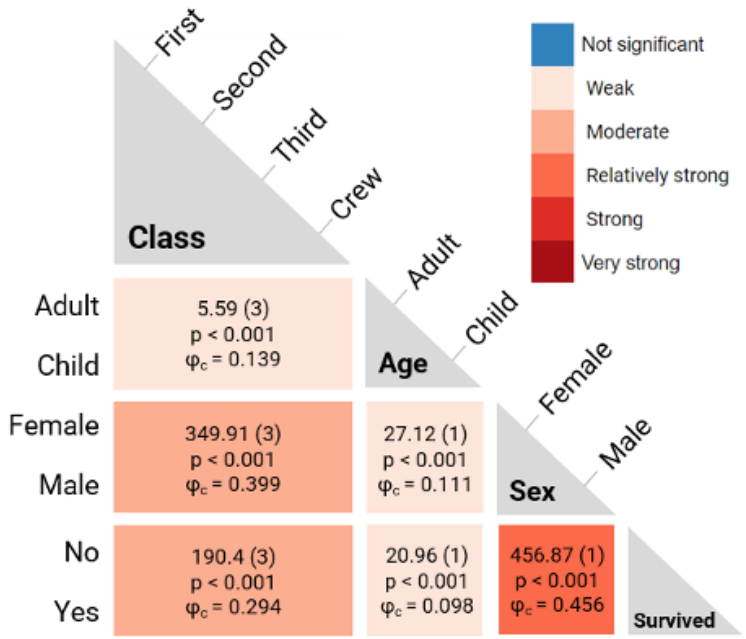


Figure 4.5: Panel-level test results for the Titanic dataset, including the Chi-square statistic, degrees of freedom, p-value and Cramér's V. Cell colour is proportional to the strength of the association.

test results are shown for panels that satisfy the two remaining conditions, namely that categories within each variable are mutually exclusive, and that expected frequencies exceed one in all cells and are at least five in 80% of cells.

If the user confirms independence of observations and the remaining test conditions are satisfied, the corresponding panel is coloured either red or blue, depending on the test result. Red indicates a significant result, whereas blue does not. The shade of red is proportional to the strength of the association, as measured by Cramér's V: a number between 0 and 1, with larger/darker values indicating a stronger association. For completeness, the Chi-square statistic, degrees of freedom, p-value and Cramér's V are all reported in the corresponding panel. Users can also change the significance level in the text box from its default value of 0.05; this updates the test results accordingly. The legend is interactive, such that hovering over one of the values isolates all variable pairs with the corresponding effect size.

If any of the test conditions for a pair of variables is violated, the panel is coloured grey. An error message explaining the reason why the test result was not valid is shown in the tooltip. This is helpful even for datasets where majority of the panels are grey, because it shows the user that the Chi-square test is not an appropriate technique for such data, while perhaps still revealing a handful of associations that are significant.

Embedding Chi-square test results into the plot in this way has a number of benefits: it removes the burden of manual computation (which is particularly onerous for datasets with many variables), visually reinforces correct interpretations, and enables all relevant data to be conveniently displayed in one place. Furthermore, the results in this view can be effectively coupled with the cell-level Chi-square values and Pearson residuals discussed in Section 4.5.1 (McHugh, 2013). For instance, Figure 4.5 shows that there is a relatively strong association between the variables "Sex" and "Survived", and the cell-level metrics, including Figure 4.2, suggest that this is due to more females surviving than would be expected by chance, and more men dying. This aligns with the societal norm of prioritising the rescue of "women [and children] first".

4.6 Linked Table View

The linked table view, shown in Figure 4.1, connects the heatmap with the underlying data. The user can click on a cell to see the corresponding data items in the table beneath the matrix. Upon being clicked, the cell is given a black border to show that it has been selected. By default, all variables are displayed in the table, with any unique, 'ID-like' variables being shown to the immediate left of all others. For example, in Figure 4.1, the user has clicked on a cell representing female crew members and the corresponding records, including people's names, are displayed in the table. The user can navigate with the scroll bar or expand the table to view records that are not currently visible. Additionally, the table columns can be hidden or reordered. The number of rows in the table matches the cell's observed frequency; there are 23 female crew members and thus 23 records in the table. This feature is most useful if the dataset contains one or more 'ID-like' columns, and if a large proportion of cells have relatively low counts, so that the information presented in the table can be readily absorbed.

While not currently supported, it would be possible to generate supplementary visualisations from the conditional table data. One could imagine hovering over one of the column headings to display a bar chart of the number of occurrences in the table of each category from the corresponding variable. For example, hovering over the "Sex" column would show how many female crew members survived, $P(Yes \mid Female \cap Crew)$, and how many died, $P(No \mid Female \cap Crew)$.

4.7 Selection Menu

For complex datasets with a large number of categories and variables, it may not be feasible to view everything at once. A better approach might be to break down the dataset into smaller units of interest, and rotate among these. As shown in Figure 4.1 and Figure 4.6, the *Selection Menu* consists of an expandable list of checkboxes, with variables at the top level (whose cardinalities are indicated in parentheses) and categories nested inside them. The user can click on the checkboxes to show or hide variables in the matrix. Variables that are currently visible are shown with a black tick, and those filtered out with a blue box. It is also possible to show, hide *or exclude* individual categories, with three clicks required to return to each state. Excluded categories are shown with a red cross icon. The distinction between hiding and excluding a category is that the former simply removes it from the display (without affecting the rest of the matrix), whereas the latter removes all data items associated with that category, likely resulting in changes to other panels. For instance, removing children from the Titanic dataset would update *all* panels to only include information about adults, effectively resulting in a conditional query: $P(X \cap Y \mid Adult)$, where X and Y are the variables on either axis. The advantage of having the checkboxes is that users can re-select categories that they previously hid or excluded. "Select all" and "Clear" links are also available to expedite variable and category selections.

Filtering via the *Selection Menu* is primarily intended for analysing datasets with dozens of variables and/or categories (e.g., census data). This feature is not necessary for relatively simple datasets like the Titanic dataset, where all of the variables and categories can be visualised at once. However, even in such cases, the ability to exclude variables is helpful for visualising conditional queries. Overall, the addition of this menu increases the scalability of the heatmap matrix technique, albeit by requiring the user to work with manually defined subsets. Similar functionality could be added to other pairwise techniques for categorical data, such as the Mosaic Matrix (Friendly, 1999).

Summary statistics at the top of the menu indicate the proportion of variables and categories visible at any given time, as well as how many data items have been excluded. It might also be helpful to report statistics such as "Categories shown for *selected* variables" (to exclude variables that the user deems irrelevant) or the "(Average) number of categories per variable". These could all be visualised as stacked bar charts or histograms, rather than being given as text labels.

The *Selection Menu* also provides a mechanism for manually re-ordering the data in the heatmap, by dragging-and-dropping the labels. The user can move variables, as well as the categories within them. The order of rows in the matrix mirrors the top-to-bottom ordering in the list of checkboxes, and columns are ordered in the same way from left to right. For instance, moving "Class" beneath "Age" in Figure 4.1 would make "Age" the top/left-most box, and dragging the category "Crew" above "First" would make "Crew" the top/left-most category within the "Class" portion of the matrix.

A final operation supported by the menu is merging existing categories. The user can use Shift-click and Ctrl-click to select multiple categories for a particular variable, then right click on one of them to merge those categories. They are then given the option to rename the newly formed category. The number of categories is automatically reduced to reflect the number that remain, with undo and redo functionality supported in case the user makes a mistake or wishes to revisit a previous state.

4.8 Covid Directives Dataset

All the examples given so far concern the Titanic dataset. Figure 4.6 shows a more complex example of a heatmap matrix, illustrating a linguistics dataset comprising ten categorical variables (Burnette and Calude, 2022). The data consist of directives used in tweets featuring the hashtag "#covid19nz" (e.g., "Stay home!"). This dataset was compiled to examine pragmatic and syntactic variables in relation to the stance of directives towards COVID-19 government measures in New Zealand, during the first nation-wide lockdown. The user has hidden four variables from the display, and removed two categories from "Stance": "None" and "Unclear". This has resulted in 88 of the 754 directives being filtered out of the heatmap. The matrix view shows that there is one dominant pair of categories (or 'flavour' of directive) in each and every panel. For instance, the panels for "Stance" and "Politeness" show that those in most agreement with the status quo ("pro") were also least concerned to mitigate their directive with polite markers ("no redress"). "Stance" and "Verb" exhibit greater variation than any other pair of variables, with main verbs and modal verbs being relatively common across all three stance categories.

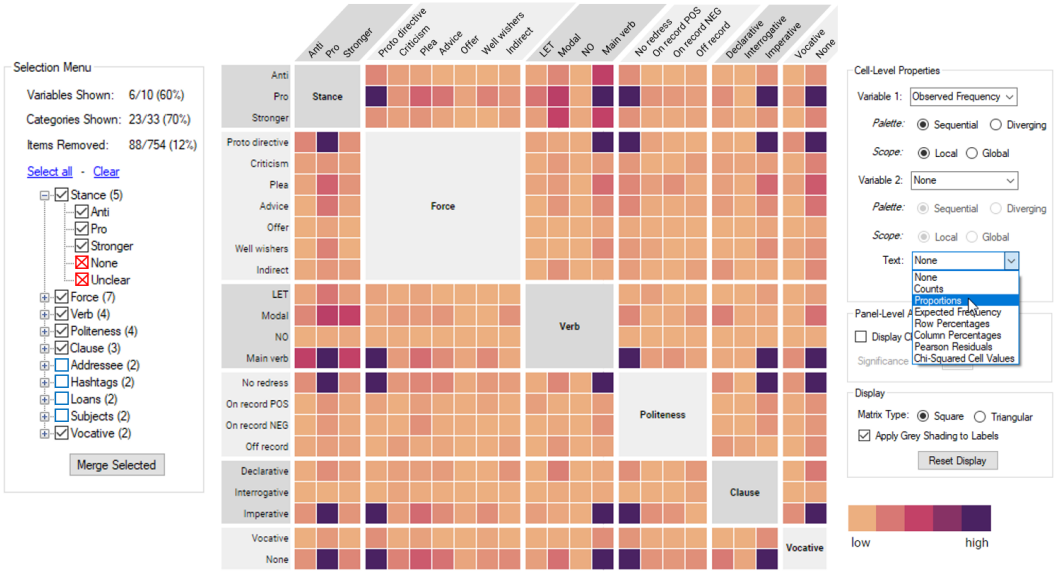


Figure 4.6: A more complex example of a heatmap matrix, showing information about Covid directives on Twitter. Some variables have been hidden, and two categories have been excluded from the data, as indicated in the *Selection Menu* on the left-hand side.

4.9 Limitations

The *Heatmap Matrix Explorer* has some limitations that need to be acknowledged. First, it does not incorporate several of the useful interactive features described by Rocha and da Silva (2022), such as automated methods for sorting the matrix, which would be useful for revealing structural patterns, or the ability to bin continuous values. Second, a lot of the design decisions are based on the authors' subjective preferences and require more comprehensive user testing. For instance, the thin white borders around cells might actually be a distraction for perceiving patterns within and between different panels. Third, displaying cell Chi-square values from the drop-down menu for an invalid Chi-square test may be problematic, and there is currently nothing to safeguard against this. Furthermore, the Chi-square test and Cramér's V are not well suited to ordinal data, as they do not consider ordering information. The datasets in this paper contain mostly nominal variables, but a Spearman correlation or Kendall's Tau would be more appropriate for panels involving strictly ordinal data. While there are, in fact, several alternative methods for analysing categorical data, the bigger picture is that such tests can be effectively embedded into visualisations to aid the viewer's understanding.

4.10 Conclusions and Future Work

This paper has proposed a structured set of extensions for augmenting the heatmap matrix, which are realised in an empirical prototype called the *Heatmap Matrix Explorer*. These extensions improve the readability, versatility and scalability of the heatmap matrix technique. The revised design removes non-bivariate cells, re-positions variable labels, removes dense grid lines and has a white background. Interactive drop-down menus allow the user to colour and label cells according to several metrics, including row percentages and expected frequencies. The high-level overview for the Chi-square test helps the viewer to quickly detect patterns and establish which variables have the strongest associations. Examining these findings in relation to cell-level metrics like Pearson residuals and individual Chi-square values can then provide more detailed information about specific cells driving the association. The *Linked Table View* provides a direct and convenient link to individual records, and the *Selection Menu* enables exploration of more complex datasets than was previously possible, by allowing controlled yet flexible filtering. Overall, these extensions provide greater insight into the relationships between categorical variables, by encouraging the user to explore the data from a range of perspectives, and empowering them to uncover more complex patterns in the process.

Future work could centre around turning the empirical prototype into a web-based tool that allows users to visualise their own categorical datasets, and conducting in-depth user testing. Two further avenues of inquiry are dealing with missing values, which may differ across variables, and supporting nested heatmaps for hierarchical categorical data.

4.11 Postscript

This chapter has shown how the readability, functionality and scalability of an existing visualisation technique can be improved to more effectively support the analysis of several categorical variables. More specifically, our empirical prototype—the Heatmap Matrix Explorer—helps users to uncover more detailed information about potential associations in a categorical dataset, and is more scalable than other bivariate techniques, such as the Mosaic Matrix. We will utilise the Heatmap Matrix Explorer in one of our case studies in Part III (Chapter 8) to further demonstrate its value.

4.12 References

Agresti, A. (2013). *Categorical data analysis*. John Wiley & Sons, 3 edition.

Alsallakh, B., Aigner, W., Miksch, S., and Gröller, M. E. (2012). Reinventing the Contingency Wheel: Scalable visual analytics of large categorical data. *IEEE Trans. Vis. Comput. Graphics*, 18(12):2849–2858.

Alsallakh, B., Gröller, M. E., Miksch, S., and Suntinger, M. (2011). Contingency Wheel: Visual analysis of large contingency tables. In *EuroVA@EuroVis*.

Becker, R. A., Cleveland, W. S., and Shyu, M.-J. (1996). The visual design and control of trellis display. *J. Comput. Graph. Stat.*, 5(2):123–155.

Burnette, J. and Calude, A. S. (2022). Wake up New Zealand! Directives, politeness and stance in Twitter #covid19nz posts. *J. Pragmat.*, 196:6–23.

Carr, D. B., Littlefield, R. J., Nicholson, W., and Littlefield, J. (1987). Scatterplot matrix techniques for large n. *J. Am. Stat. Assoc.*, 82(398):424–436.

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *J. Stat. Educ.*, 3(3).

Dougherty, J., Kohavi, R., and Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning proceedings 1995*, pages 194–202. Elsevier.

Elmqvist, N., Dragicevic, P., and Fekete, J.-D. (2008). Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Trans. Vis. Comput. Graphics*, 14(6):1539–1148.

Emerson, J. W., Green, W. A., Schloerke, B., Crowley, J., Cook, D., Hofmann, H., and Wickham, H. (2013). The generalized pairs plot. *J. Comput. Graph. Stat.*, 22(1):79–91.

Fernstad, S. J. and Johansson, J. (2011). A task based performance evaluation of visualization approaches for categorical data analysis. In *2011 15th International Conference on Information Visualisation*, pages 80–89. IEEE.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., and Hullman, J. (2021). The science of visual data communication: What works. *Psychol. Sci. Public Interest*, 22(3):110–161.

Friendly, M. (1992). Graphical methods for categorical data. *Proceedings of SAS SUGI*, 17:190–200.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *J. Am. Stat. Assoc.*, 89(425):190–200.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *J. Comput. Graph. Stat.*, 8(3):373–395.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices.

Am. Stat., 56(4):316–324.

Gehlenborg, N. and Wong, B. (2012). Heat maps. *Nat. Methods*, 9(3):213.

Greenacre, M. (2017). *Correspondence analysis in practice*. CRC press.

Hartigan, J. A. and Kleiner, B. (1984). A mosaic of television ratings. *Am. Stat.*, 38(1):32–35.

Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika*, 51:11–26.

Hofmann, H. and Vendettuoli, M. (2013). Common angle plots as perception-true visualizations of categorical associations. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2297–2305.

Im, J.-F., McGuffin, M. J., and Leung, R. (2013). GPLOM: The generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Trans. Vis. Comput. Graphics*, 19(12):2606–2614.

Koffka, K. (1935). *Principles of Gestalt Psychology*. Lund Humphries.

Kosara, R., Bendix, F., and Hauser, H. (2006). Parallel Sets: Interactive exploration and visual analysis of categorical data. *IEEE Trans. Vis. Comput. Graphics*, 12(4):558–568.

McHugh, M. L. (2013). The chi-square test of independence. *Biochemia medica*, 23(2):143–149.

Meyer, D., Zeileis, A., Hornik, K., and Leisch, F. (2003). Visualizing independence using extended association plots. *Proceedings of DSC 2003*.

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Reza, R. M. and Watson, B. A. (2019). Hi-D maps: An interactive visualization technique for multi-dimensional categorical data. In *2019 IEEE visualization conference (VIS)*, pages 216–220. IEEE.

Rocha, M. and da Silva, C. G. (2022). Heatmap matrix: Using reordering, discretization and filtering resources to assist multidimensional data analysis. `https://doi.org/10.13140/RG.2.2.36619.57126`.

Rocha, M. M. N. and da Silva, C. G. (2018). Heatmap matrix: a multidimensional data visualization technique. In *Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI)*.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE.

Waskom, M. L. (2021). seaborn: statistical data visualization. *J. Open Source Softw.*, 6(60):3021–3024.

Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J., et al. (2017). Package 'corrplot'. *Statistician*, 56:316–324.

Wilkinson, L., Anand, A., and Grossman, R. (2005). Graph-theoretic scagnostics. In *2005 IEEE Symposium on Information Visualization*, pages 157–164. IEEE.

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *J. Comput. Graph. Stat.*, 16(3):507–525.

Chapter 5

MultiCat: A Visualisation Technique for Multidimensional Categorical Data

As discussed in Chapter 3, there are clear limitations of existing techniques for visualising multivariate categorical data. In this chapter, we introduce *MultiCat*, a novel interactive technique for exploring a large number of categorical variables simultaneously. In contrast to the Heatmap Matrix discussed in the previous chapter, which breaks down categories into pairwise relationships, MultiCat supports analysis of higher-order (multivariate) relationships. It is therefore well-equipped to provide insights that are distinct from—but complementary to—those revealed by the Heatmap Matrix Explorer. To facilitate broader application of MultiCat, we have made our code publicly available, together with an online demonstration of the tool.

Manuscript Details

This manuscript, co-authored with Mark Apperley and David Bainbridge, has not been published at the time of submission of this thesis.

Abstract

Exploring relationships among several variables is an important analysis task when dealing with multidimensional categorical data. A key challenge in visualising such data lies in ensuring that each variable and its categories can be clearly distinguished, especially when these grow in number and complexity. This paper presents MultiCat, an interactive visualisation technique for

analysing categorical datasets comprising roughly 3–20 variables. MultiCat uses a familiar, spreadsheet-like layout to represent both nominal and ordinal variables in frequency form. It incorporates several interactive features, including column-wise sorting, dynamic queries and filtering, and allows rapid calculation of *a priori* and conditional probabilities. MultiCat offers several advantages over existing techniques, including: (1) enhanced clarity in extracting high-dimensional relationships and comparing their frequencies; (2) a non-hierarchical default layout that promotes user-driven analysis; and (3) a structured visual overview of the relative contribution of each category. We validate MultiCat by reporting on the promising results and outcomes of a small-scale usability study. A prototype of MultiCat is available at `https://dgt12.github.io/multicat/`.

5.1 Introduction

Understanding relationships among variables is an important analysis task, whether those variables are categorical or continuous. Categorical variables are frequently encountered in real-world datasets, ranging across such varied domains as behavioural and social sciences, public health, biomedical science, education and marketing (Agresti, 2012). For example, categories can be used to represent patient treatment outcomes (no improvement, some improvement, marked improvement), survey responses (strongly disagree to strongly agree) or customer brand preferences (Brand X, Brand Y, Brand Z). However, despite their prevalence, few visualisation techniques support the analysis of more than three categorical variables at the same time.

A significant challenge in visualising categorical data lies in ensuring that each variable and its categories can be clearly distinguished, regardless of how many there are. In addition, given that nominal variables do not have an intrinsic order, it is difficult to know how best to arrange them. Existing methods for visualising multidimensional categorical data either do not scale well (Hartigan and Kleiner, 1981; Greenacre, 2017; Tenenhaus and Young, 1985; Reza and Watson, 2019) or fail to consider relationships among all variables simultaneously. They typically break down the data into more restricted views, such as pairwise relationships (Rocha and da Silva, 2018; Trye et al., 2023; Friendly, 1999; Im et al., 2013; Greenacre, 2017; Tenenhaus and Young, 1985), or impose a hierarchy of variables (Kosara et al., 2006; Hartigan and Kleiner, 1981; Kolatch and Weinstein, 2001), which affects what insights can be seen. Moreover, these techniques often lack code-free, user-friendly interfaces,

limiting their accessibility to a broader audience.

Recognising this gap, we adopt a technique-driven approach (Sedlmair et al., 2012) to design and validate MultiCat, a novel method for visualising multidimensional categorical data. MultiCat allows users to generate new insights and hypotheses about the interplay of categories across as many as 20 variables, by focusing on both individual categories and their higher-dimensional relationships. This is accomplished by using a tabular visualisation of the data in frequency form, coupled with a sidebar that comprises multiple linked bar charts. MultiCat also serves as an interactive probability calculator, helping users to compute and reason about a wide range of *a priori* and conditional probabilities. We validate MultiCat with a small-scale usability study, from which we have used feedback and observations to improve our prototype. MultiCat is generalisable across datasets and domains, and is therefore of interest to anyone who works with categorical data, including social scientists, business analysts and marketing experts.

Throughout the paper, we use the **Titanic dataset** (Dawson, 1995) as our primary example. This dataset, compiled by Robert Dawson in 1995, details socio-historical information about the people aboard the RMS *Titanic* when it tragically sank in 1912. It has been visualised extensively in the context of categorical data analysis (Symanzik et al., 2019). The dataset contains 2,201 observations (people) and comprises four categorical variables: Class (first, second, third and crew), Sex (male, female), Age (child, adult) and Fate (survived, died). This last variable has been renamed from Survived (yes, no) to provide more descriptive category names. Given its modest size and absence of missing values, the Titanic dataset is well-suited to introducing the MultiCat technique. At the same time, its widespread use enables a direct comparison with other methods (see, for instance, Figures 5.1-5.3). When describing MultiCat, we emphasise design features that make it suitable for handling more complex datasets, drawing on other examples where necessary.

The structure of the paper is as follows: We begin by stating contributions, discussing key terminology and surveying related work. Based on the capabilities and limitations of existing techniques, we outline a set of design requirements that informed the development of MultiCat. We then introduce the MultiCat technique by focusing on its spreadsheet view and sidebar, before comparing it with an earlier design. Next, we delve into MultiCat's interactive features, which range from dynamic queries to sorting and filtering. Implementation details of our prototype are provided, followed by a discussion of scalability constraints. We then describe the methodology and results of a user

study aimed at identifying usability issues and gathering general feedback. A direct comparison with two other techniques is given, highlighting MultiCat's unique advantages and areas for improvement. Following this, we propose a series of extensions and enhancements for MultiCat. The paper concludes with a summary of the contributions of our research and opportunities for future work.

5.1.1 Contributions

This paper makes the following contributions:

1. The design and implementation of MultiCat, a novel visualisation technique for analysing multidimensional categorical data.
2. A small-scale usability study that sheds light on the value of this technique and highlights opportunities for further improvement.

5.1.2 Terminology

Since a variety of terms are used in the literature in relation to categorical data, we detail our adopted usage here. The term **multidimensional** is used throughout the paper to refer specifically to three or more categorical variables. We primarily refer to each entity in a dataset as a **data item**, rather than as a "record", "case" or "observation". Each discrete set of values is described as a **variable**, rather than as an "attribute" or "dimension", and the values themselves are designated as **categories** rather than "levels" or "classes". We consider a categorical variable to be either **nominal** (unordered) or **ordinal** (categories with a natural ordering), and believe it is important for a categorical visualisation tool to accommodate both types. The term **colour** is used throughout the paper to refer specifically to "hue". Finally, **cardinality** denotes the number of categories belonging to some variable, such that a high-cardinality variable has many (10 or more) categories.

5.2 Related work

We discuss related work in the context of prominent approaches and techniques for visualising multidimensional categorical data, before outlining relevant connections to interactive tables and hypergraphs.

5.2.1 Multidimensional categorical data

Over the past few decades, a number of techniques for visualising multidimensional categorical data have emerged, yet their adoption has not always been widespread (Theus, 2012). Most of these techniques are derived from contingency tables (Alsallakh et al., 2012), either by representing the cell counts directly or projecting categories into a two-dimensional plane. Following previous work (Johansson Fernstad and Johansson, 2011), we refer to these two approaches as "CatViz" and "QuantViz" methods, respectively. In general, CatViz methods are "lossless" (Dimara et al., 2017) and more effective for frequency-based tasks, whereas QuantViz methods are "lossy" and better suited to similarity-based tasks (Johansson Fernstad and Johansson, 2011). Due to limitations of space, we cover only three established techniques here, focusing on those which are most commonly cited in the literature. A far more comprehensive database of relevant techniques is available at `https://cat-vis.github.io`.

5.2.2 Interactive Mosaic Plots

Mosaic Plots (Hartigan and Kleiner, 1981) have been described as the "Swiss Army knife" of categorical data displays (Theus, 2012). These plots fall under the CatViz umbrella and are created by recursively subdividing variables along alternate axes, forming area-proportional tiles. If the tiles are neatly aligned, this means the variables are independent (Friendly, 1999). Residual-based shading of tiles is sometimes also used to visualise loglinear models (Friendly, 1994) and statistical significance of test results (Zeileis et al., 2007). Interactive Mosaic Plots are available in a variety of tools, including Mondrian (Theus, 2002), ViSta (Young and Bann, 1996) and MANET (Unwin et al., 1996), greatly enhancing their exploratory power. For instance, Mondrian allows users to switch between multiple variants of Mosaic Plots (Theus, 2012), add, remove or rotate variables, select different regions and access tooltips for each tile. Moreover, users can probe complex relationships by querying the data via linked bar charts, as shown in Figure 5.1. A major limitation of Mosaic Plots, however, is that they become increasingly difficult to read when displaying more than a handful of variables and/or categories. This leads to an increase in empty combinations and skewed tiles (Hofmann, 2006), exacerbated by low-frequency categories.

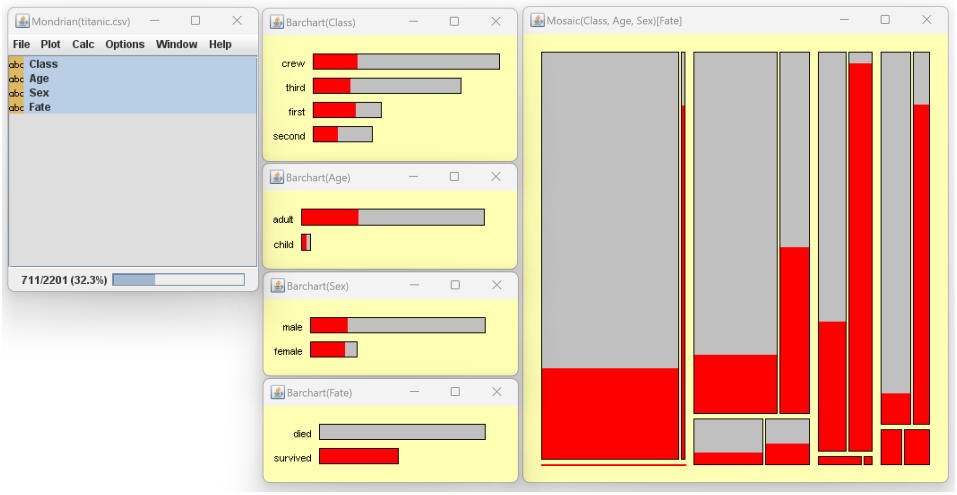


Figure 5.1: The Mondrian interface showing a Mosaic Plot with linked bar charts for the Titanic dataset (Dawson, 1995). Survivors are highlighted in red.

5.2.3 Parallel Sets

Perhaps the most scalable technique for visualising multidimensional categorical data is Parallel Sets (Kosara et al., 2006), reminiscent of Sankey Diagrams (Schmidt, 2006). Another area-proportional, CatViz technique, this method represents variables in stacked "tiers" of equal width. Associations between subsets are then shown using shaded parallelograms connecting adjacent tiers; see Figure 5.2. Parallel Sets visualisations are capable of handling 10–15 categorical variables in an interactive environment (Kosara et al., 2006), and 20–30 categories in total. While Parallel Sets supports rich interaction, including flexible reordering of variables and categories, it invariably suffers from line crossings and perceptual distortions (Hofmann and Vendettuoli, 2013). These are exacerbated by the hierarchical nature of the display. Furthermore, changing the aspect ratio of the visualisation can alter the appearance of the parallelograms, yielding skewed results. These disadvantages are partially addressed by Common Angle Plots (Hofmann and Vendettuoli, 2013) and Hammock Plots (Schonlau, 2003) but, in all cases, the order in which variables are plotted can drastically change the visualisation. Various quality metrics for evaluating Parallel Sets have been proposed, with a view to reducing visual clutter (Dennig et al., 2021; Zhang et al., 2019).

Although an academic prototype was developed for Parallel Sets, it is no longer maintained and does not appear to run on modern machines (par, 2009). However, Parallel Sets has been implemented as a reusable D3.js (Bostock et al., 2011) chart, together with the most important interactive fea-

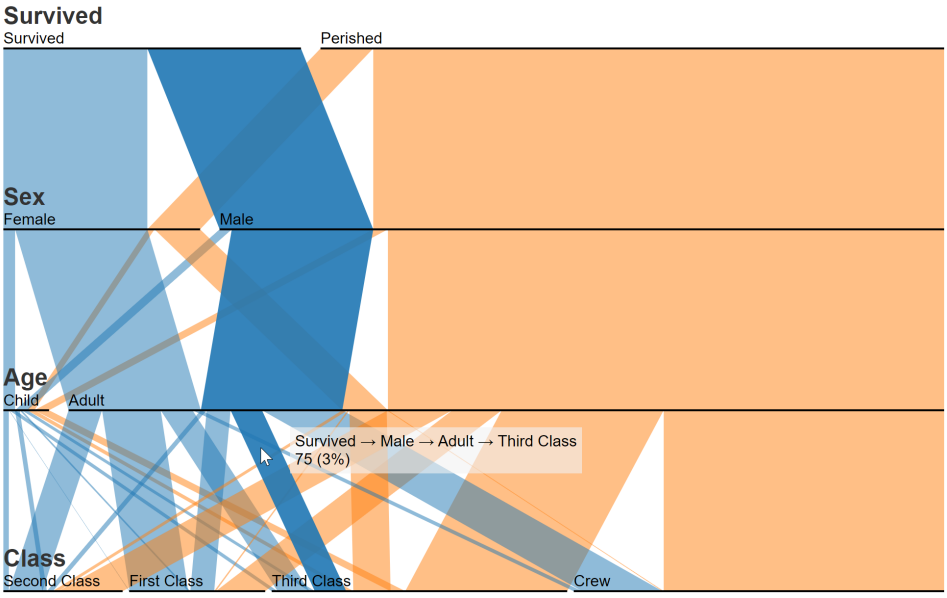


Figure 5.2: A Parallel Sets visualisation of the Titanic dataset (Davies, 2012).

tures (Davies, 2012). Related visualisation techniques, such as (Categorical) Treemaps (Kolatch and Weinstein, 2001), are also available in code-free tools like RAWGraphs (Mauri et al., 2017). Additionally, the R package *ggparallel* (Hofmann and Vendettuoli, 2013) creates static, pairwise visualisations of Parallel Sets, Hammock Plots and Common Angle Plots. Nevertheless, these alternatives do not offer the full range of features described in the original Parallel Sets papers (Bendix et al., 2005; Kosara et al., 2006), such as the ability to view histograms, merge categories or select multiple parallelograms to visualise the corresponding proportion of data. This exemplifies a broader issue endemic to the field: the divide between theoretical innovation and practical application of novel visualisation techniques.

5.2.4 Correspondence Analysis

Correspondence Analysis (CA) (Greenacre, 2017) is a widely used QuantViz method that shows associations in a two-way contingency table. The row and column categories in a table are depicted as points on a graph whose positions indicate associations between categories.

Multiple Correspondence Analysis (MCA) (Tenenhaus and Young, 1985) extends this principle to n-way tables, accommodating analyses involving more than two variables (see Figure 5.3). While MCA provides a broader scope than CA, it still focuses on pairwise relationships. MCA typically provides a visual representation of the so-called "Burt Matrix", which encodes the joint, bivariate relations between every pair of variables in a dataset (Friendly and

Meyer, 2015).

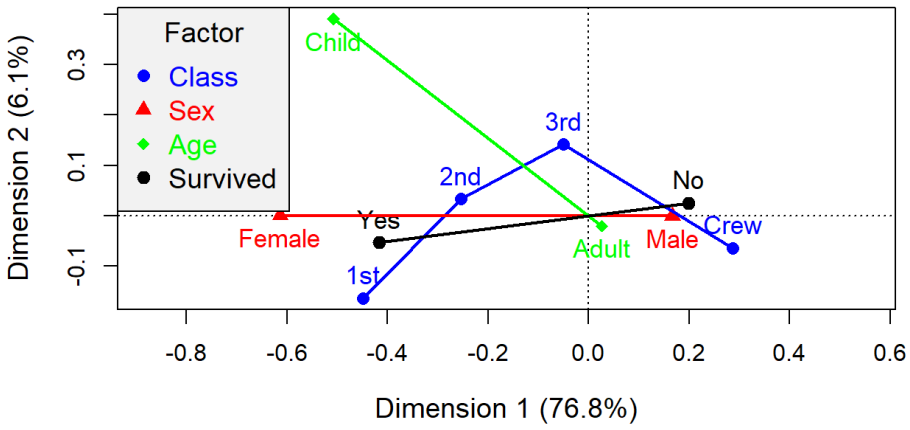


Figure 5.3: MCA Plot of the Titanic dataset (Friendly and Meyer, 2015).

While CA and MCA are useful for capturing structure in high-dimensional categorical datasets, they have a number of drawbacks. Both techniques are difficult for non-experts to interpret, they do not display frequency-related information, or convey the reasons *why* items belong to particular clusters. Furthermore, CA and MCA quickly become cluttered when the number of categories increases, since the individual category labels are usually shown next to the points themselves. When there are large numbers of variables in MCA, it is also difficult to determine which categories belong to which variables.

Motivated by the limitations of these existing solutions, we adopt a technique-driven approach in this work to design and validate a novel method for exploring and analysing multidimensional categorical data.

5.2.5 Connection to tabular data

Multidimensional categorical data lends itself to tabular representations. Tabular visualisation techniques employ a spreadsheet-like layout, where rows correspond to individual data items and columns correspond to variables. These techniques are geared towards understanding the properties of items by considering all variables simultaneously. Cells use visual channels such as position, length and colour to enhance readability and facilitate exploration of higher-level trends. The two tabular visualisation techniques most closely related to MultiCat are Taggle (Furmanova et al., 2020) and TableLens (Rao and Card, 1994), both of which support heterogeneous data (i.e., both continuous and categorical variables). However, while very powerful, these techniques are not

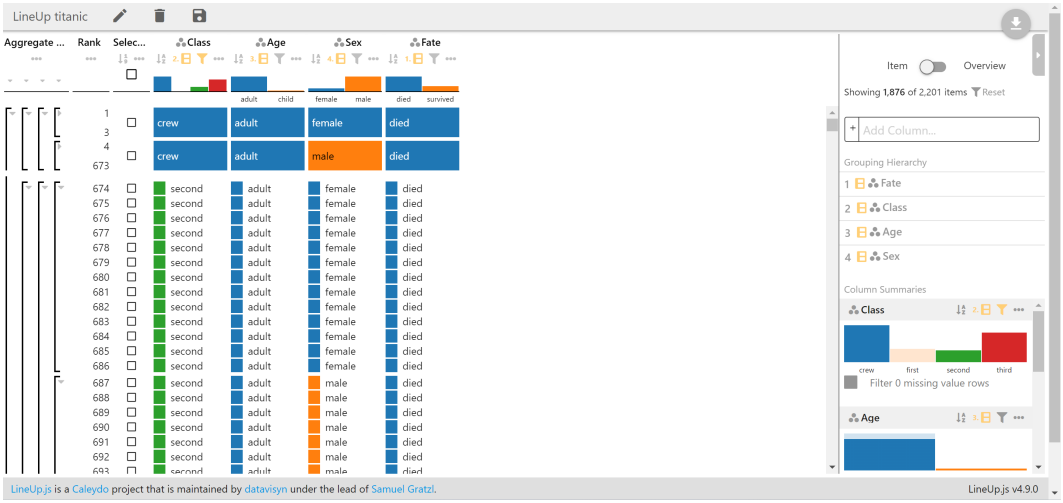


Figure 5.4: Taggle is not optimised for purely categorical data.

optimised for categorical data, as is evident in Figure 5.4. TableLens, for instance, does not support aggregation, limiting its scalability. Taggle, while offering aggregation and a height-proportional layout that reflects frequencies in its overview mode, does not provide the compactness of MultiCat's aligned bar chart encoding. Furthermore, Taggle can be confusing to navigate when visualising purely categorical data, as many of its features were not intended for such data. MultiCat provides a more focused analysis by removing extraneous features and streamlining its workflow for categorical data.

It has been convincingly argued that interactive tables are an important visualisation technique in their own right (Bartram et al., 2021). Spreadsheet applications like Microsoft Excel and Google Sheets play a critical role in helping users to make sense of data, incorporating powerful features like sorting and filtering. However, at the same time, these applications lack custom interaction, provide limited support for visual encoding of cells, make assumptions about how the data should be handed (e.g., sorting categorical values alphabetically rather than by frequency) and do not offer multiple coordinated views (Gratzl et al., 2013). MultiCat seeks to overcome these limitations from a categorical data perspective, while preserving the essence and functionality of an interactive table.

5.2.6 Connection to hypergraphs

Finally, we note that multidimensional categorical data can be accurately represented in a *hypergraph* (Fischer et al., 2021). This structure is an extension of a traditional graph: the key difference is that its edges, termed *hyperedges*, can connect any number of vertices. A hyperedge is therefore equivalent to a

set. Our initial design for MultiCat, depicted in Figure 5.9, was inspired by PAOHVis (Valdivia et al., 2021), which employs a matrix layout, displaying the vertices of a hypergraph as circular nodes on one axis, and hyperedges as connecting lines on the other. While experimenting with this technique, we realised that the categories in a dataset could be treated as vertices and their orthogonal combinations as hyperedges. Previous work linking hypergraphs with multidimensional categorical data has done the opposite, representing data items as vertices and categories as hyperedges (Nguyen and Mamitsuka, 2020). Although the final design of MultiCat differs significantly from that of PAOHVis, this early inspiration was crucial, and we believe that modelling categorical data as hypergraphs may be fruitful in a variety of other contexts.

5.3 Design requirements

We developed the following set of design requirements for MultiCat by assessing the capabilities and limitations of the above techniques. From the outset, we decided to adopt an aggregation-based approach that provided a direct representation of category counts.

R1: Aggregate categories. Provide a compact visual representation of the data in frequency form, ensuring that the categories within each combination/aggregate are easily readable.

R2: Show category distributions. Include univariate summaries for each variable that can be readily compared. Users should be able to extract absolute values and marginal frequencies for each category.

R3: Support multiple variables. Allow the user to visualise 3–20 categorical variables. Variables should be treated as equally as possible, and changing their order should not drastically alter the display. Additionally, the layout used should be relatively independent of the number of data items.

R4: Support high-cardinality variables. Ensure that the technique can handle variables with potentially large numbers of categories, while accentuating the most important/frequent categories within each variable. The cardinality of variables may differ considerably, but most variables will be expected to have between two and ten categories.

R5: Handle ordinal variables. Ordinal variables should also be supported, and they should be visually distinct from nominal ones. For ordinal variables, the inherent order of categories should be apparent in the visualisation.

R6: Allow interactive refinement and visual feedback. Users should

be able to dynamically add and remove variables, and to select/query different subsets of categories. The percentage of selected data should always be visible, and the display should update immediately when the user interacts with it.

R7: Incorporate filtering. The interface should allow users to filter the data and compute conditional probabilities from the resultant subsets.

R8: Incorporate sorting. Users should be able to efficiently sort categorical and numeric values. It should be possible to sort by multiple columns in order to break ties at higher levels.

R9: Use a minimalist design. The interface should avoid unnecessary features that detract from the above requirements.

These nine requirements have guided the development and evaluation of our proposed technique for visualising multidimensional categorical data.

5.3.1 Assumptions

We make the following assumptions about the data to be analysed within MultiCat:

1. The input dataset contains only nominal and ordinal variables.
2. Categories belonging to the same variable are mutually exclusive.
3. Categories belonging to different variables are not necessarily independent.
4. Any missing values are coded as "Unknown".

5.4 The MultiCat technique

In this section, we describe the MultiCat technique in detail and justify our design decisions with reference to visualisation theory and existing tools. We have endeavoured to use perceptually efficient visual encodings in our design, but the complexity of the data meant there were a number of trade-offs involved. Figure 5.5 shows the MultiCat interface with the Titanic dataset (Dawson, 1995) loaded in. MultiCat consists of two coordinated views: a spreadsheet view on the left, which is the main display, and a sidebar on the right. The spreadsheet view shows distinct combinations of orthogonal categories (rows) associated with a chosen set of variables (columns). The sidebar, on the other hand, displays information about individual categories and affords an intuitive means of selecting and filtering different subsets of the data. We provide more detail about the layout of each of these components, before addressing their interactive capabilities. The descriptions given here reflect our final proto-

Figure 5.5: MultiCat visualisation of the Titanic dataset (Dawson, 1995). The spreadsheet view on the left shows every observed combination of categories, aggregated and sorted by frequency. Positive (blue) residuals and negative (red) residuals indicate over- and under-represented combinations, respectively. The sidebar on the right summarises univariate category distributions, with categories grouped by variable and ordered by frequency.

type, `https://dgt12.github.io/multicat/`, which differs slightly from the version used by participants in our formative user study.

5.4.1 Spreadsheet view

The spreadsheet view in MultiCat provides a compact visual representation of categorical data in *frequency form* (Friendly and Meyer, 2015), displaying each combination of orthogonal categories for the selected variables. Rows represent distinct category combinations and columns represent variables. In the two right-most columns, frequency and Pearson residual values are shown as embedded bar charts (Gratzl et al., 2013) in a similar manner to UpSet's "Cardinality" and "Deviation" metrics (Lex et al., 2014). This arrangement aggregates items with shared characteristics, fulfilling design requirement R1. For instance, the top rows in Figure 5.5 indicate that the largest groups of

people on board the Titanic were 670 male adult crew members who tragically died and 387 male adult third-class passengers who suffered the same fate. The categories within each combination are represented by colour-coded "stickers" with text labels. For nominal data, the colour assignment is based on the frequency ranking within each variable: the most frequent category is blue, followed by green, and so on, with categories beyond fifth position being coloured grey. To maintain uniform column widths, variable and category labels are truncated as necessary, with tooltips displaying the full names.

MultiCat's category stickers enhance perceptual processing by combining colour and text within a single component. In the visualisation literature, the closest counterparts to these stickers are Taggle's (Furmanova et al., 2020) five "Item Visualization" options for categorical data. However, these approaches either separate colour from text, decreasing visual immediacy (see Figure 5.6, left column), or align category icons horizontally, which does not accommodate fixed text labels or high-cardinality variables (see Figure 5.6, right column). MultiCat addresses these issues by effectively balancing perceptual recognition with spatial efficiency.

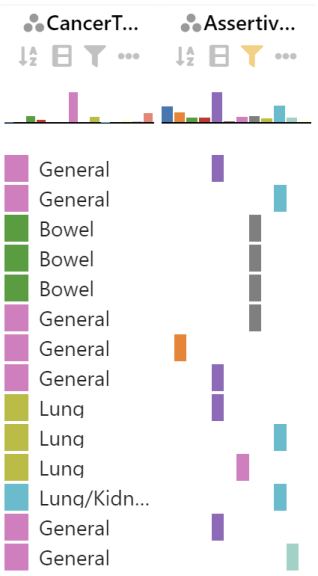


Figure 5.6: Taggle's "Color & Label" encoding (left column) and "Matrix" encoding (right column) are two alternatives to MultiCat's "Sticker" approach.

The right-most column in the spreadsheet view represents each combination's Pearson residual (Friendly, 1994) in a diverging bar chart. This measure shows the deviation of a combination's observed frequency from its expected frequency, assuming mutual independence between categories. Akin to UpSet's (Lex et al., 2014) "Deviation" metric, combinations occurring more or less frequently than expected are represented by blue (positive) or red (neg-

Figure 5.7: Sorting the Titanic data by residuals confirms expected patterns: male survivors were under-represented, whereas female survivors were over-represented.

ative) bars, respectively. As illustrated in Figure 5.7, the smallest negative residuals for the Titanic dataset are linked to deceased female adults and surviving male adults, while the largest positive residuals are associated with female survivors.

Combinations in the spreadsheet view are initially sorted by descending frequency, not by the categories themselves. This is demonstrated by the decreasing size of yellow bars in Figure 5.5. Sorting in this way ensures that variables are treated relatively equally, aligning with design requirement R3. More importantly, it simplifies gaining an overview of the distribution of category combinations, and facilitates identification of the most and least frequent aggregates. The least frequent combinations reveal anomalies in the data, such as one girl in first class who survived the Titanic disaster and three female crew members who did not (see bottom rows of Figure 5.5). Crucially, the spreadsheet view leverages users' familiarity with interactive tables, harnessing their depth of meaning and structural benefits, as detailed by Bartram et al. (2021).

5.4.2 Sidebar

The sidebar, positioned on the right-hand side of the display, offers a succinct summary of category distributions with respect to the current filter, satisfying design requirement R2. It features a series of horizontal bar charts, where cat-

egories are grouped by variable and sorted by frequency for nominal variables, or inherent order for ordinal variables. Each variable group includes a heading with the variable name and a checkbox, followed by individual category labels with their own checkboxes and bars. Category labels are truncated if they do not fit within the allocated space. All category bars share a common baseline and scale, enabling direct comparisons within and across variables. When no filter has been applied, as in Figure 5.5, the category bar lengths correspond to the marginal distribution of each variable. For instance, the sidebar in this figure highlights that the crew accounted for a surprisingly large proportion of people on board the Titanic, that only a small proportion of passengers were children or female, and that roughly twice as many people died as survived. The sidebar also serves as a quick reference for identifying the number of categories per variable (based on the number of bars) and determining the sequence in which colours are assigned to nominal categories (blue first, then green, etc.). This in turn helps the user to establish the relative rank of the categories within each variable when inspecting the combinations in the spreadsheet view.

The sidebar interacts with the spreadsheet view in simple yet powerful ways, as explained in the section on Interaction. At the top of the sidebar, four statistics provide useful context about the current state of the display. These statistics are expressed both as absolute values and percentages. They relate to items that are currently selected ("Selected items"), items that are currently visible but not necessarily selected ("Items considered"), distinct combinations that are currently selected ("'Selected rows"), and active variables ("Variables shown"). The "Selected items" statistic is considered the most important, and as such, it also has a yellow bar chart representing its value. The highlighted combination frequencies in the spreadsheet view necessarily sum to the number of selected items in the sidebar. There are two buttons at the bottom of the sidebar, including a "Reset" button which provides a convenient means of returning to the original display.

Our sidebar is inspired by Taggle's "Data Selection Panel" (Furmanova et al., 2020), but we have made several changes to optimise the readability and scalability of categorical data. Both sidebars display category distributions and enable direct category selection by clicking on the bars. However, as Taggle's use of vertical bars in a fixed space can lead to overcrowding with high-cardinality variables, in MultiCat we employ horizontal bars of uniform height. This maintains readability even for variables with many categories, providing a scrollbar in the situation where not all bars are visible at once.

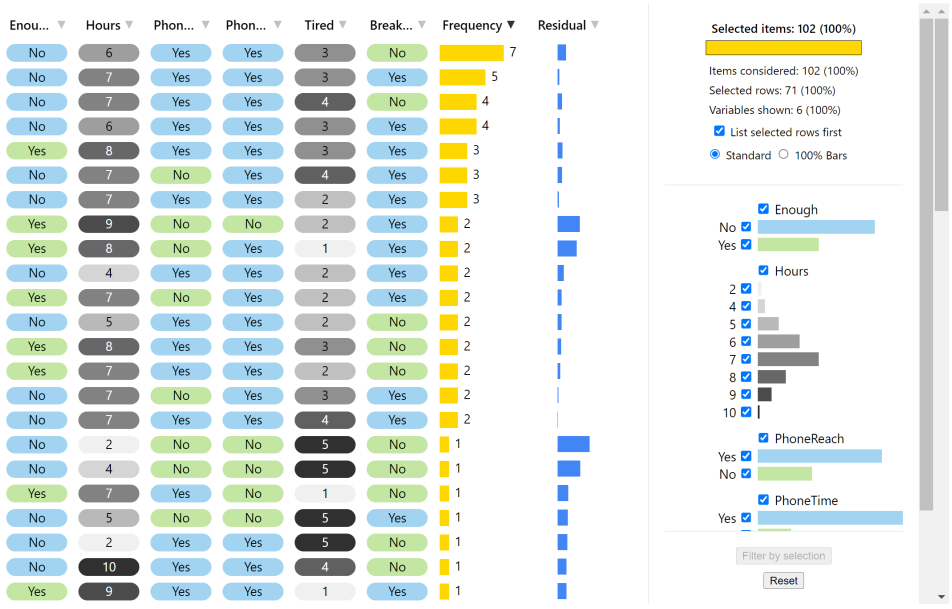


Figure 5.8: MultiCat uses greyscale values for ordinal variables. This dataset about people's sleeping habits (Lomuscio, 2020) comprises two ordinal variables and four nominal ones. Ordinal categories are displayed in their inherent order in the sidebar.

Another point of difference is that MultiCat respects the inherent order of ordinal variables, as discussed below.

5.4.3 Ordinal variables

As per design requirement R5, MultiCat can handle ordinal variables as well as nominal ones. Variables in the input dataset whose category names begin with Arabic numerals (e.g., "1 Small", "2 Medium", "3 Large") are treated as ordinal. There are two key differences regarding the appearance and behaviour of ordinal variables, which are illustrated in Figure 5.8. Firstly, these variables are depicted using greyscale values instead of hue, following Mackinlay's recommendation for a more precise visual encoding (Mackinlay, 1986). This design choice not only aligns with best practice but also ensures that ordinal variables are instantly distinguishable from nominal ones, allowing users to quickly gauge the number of variables of each type. Categories beginning with larger numbers are represented by progressively darker shades, with white text being used on darker backgrounds to ensure that category stickers remain readable. Secondly, the sorting of ordinal variables in MultiCat maintains their inherent sequential order, rather than applying a frequency-based ranking. This natural ordering is also used for categories in the sidebar. A limitation of the greyscale mapping is its reduced effectiveness in differentiating between

selected categories with lighter shades and non-selected categories with darker shades, due to the interplay of transparency with saturation (Munzner, 2014). However, the category checkboxes in the sidebar and partial transparency of entire rows (including nominal variables) allow the user to ascertain which categories have and have not been selected.

5.4.4 Colour coding

Colour coding is an effective means of displaying category information (Ware, 2019), with colour differences being more readily perceived than shape differences (Wolfe and Horowitz, 2017). However, the number of colours should be limited to between five and ten to ensure they can be rapidly distinguished (Healey, 1996; Ware, 2019). With this in mind, MultiCat employs a maximum of six colours for nominal variables, using grey for all categories beyond the fifth most frequent one. While this makes less frequent categories harder to differentiate, we consider this to be a good compromise since these categories are generally less important and tend to be dispersed across fewer distinct combinations (rows). This conservative use of colour is also in keeping with design requirement R9. The six colours chosen for MultiCat's qualitative palette were inspired by Google Sheets' drop-down presets, which are well-balanced, have similar intensity and are suitable for reading black text. Repeating colours across variables does make it harder to identify related information in each view, yet this approach is preferable to assigning a unique hue per category, which would quickly result in a palette of indistinguishable shades. It also means that the category rankings are shown consistently, as explained below.

MultiCat's colour usage aims to reduce the viewer's cognitive load by leveraging preattentive processing (Ware, 2019). Although the visualisation is comprehensible in greyscale due to its text labels and interactivity, colour enhances usability by: (1) distinguishing categories within variables; (2) linking categories across views; and (3) indicating their frequency ranking or natural order. Given that hue does not have an inherent perceptual ordering (Munzner, 2014; Muth, 2021), MultiCat prioritises this first point over the third one. Yet, because the sidebar initially shows the order in which colours are assigned, this enables users to discern patterns based on the colour of stickers within each row. This accords with the Pearson residuals in the right-most column. For example, in visualisations containing only nominal variables, the most frequent combinations would be expected to have predominantly blue stickers, as these represent the most frequent categories. Indeed, the most frequent

combination in Figure 5.5 features only categories with blue stickers. It is also apparent when relatively infrequent categories occur within a relatively frequent combination (for example, the 13 girls in second class who survived). The distribution of colours within combinations therefore aids in confirming expected trends and identifying unexpected patterns, further strengthening the use of colour in MultiCat.

5.5 Comparison with earlier design

An early version of MultiCat was proposed in a poster paper (Trye, 2022). Inspired by set-based techniques such as PAOHVis (Valdivia et al., 2021) and UpSet (Lex et al., 2014), rather than tabular approaches like Taggle (Furmanova et al., 2020), the original design stemmed from the observation that high-dimensional combinations of categories can be represented as hyperedges in a hypergraph. It is useful to compare our new design with this older one, reproduced in Figure 5.9, to show how and why it has evolved over time. This provides clear evidence of our iterative design process, while also helping to illustrate the rationale behind specific choices. We identify and explain six key changes below.

Firstly, we have transposed the matrix and re-positioned the sidebar. Previously, rows represented categories and columns represented combinations. Flipping this orientation helps to improve scalability since a multidimensional categorical dataset will typically contain a larger number of distinct combinations than variables, and it is easier to scroll vertically than horizontally. This also better aligns with a spreadsheet layout, where columns typically represent variables, rather than items or groups of items.

Secondly, text labels have been embedded within each node, resulting in the coloured "stickers" described in the previous section. This change was motivated by the need for a more readable representation. In the original design, the smaller nodes made the layout more compact but ultimately much harder to decipher, particularly in non-interactive settings. This is because each node had to be decoded by manually tracking its position along both axes to find the corresponding label, which required a much higher cognitive load than reading category labels directly.

Thirdly, we removed the connecting lines between the categories in each combination. These were superfluous and detracted from the interactive spreadsheet metaphor. For instance, if horizontal lines were present in our new design, this might deceive users into thinking that sorting happens with respect

to entire combinations rather than individual columns.

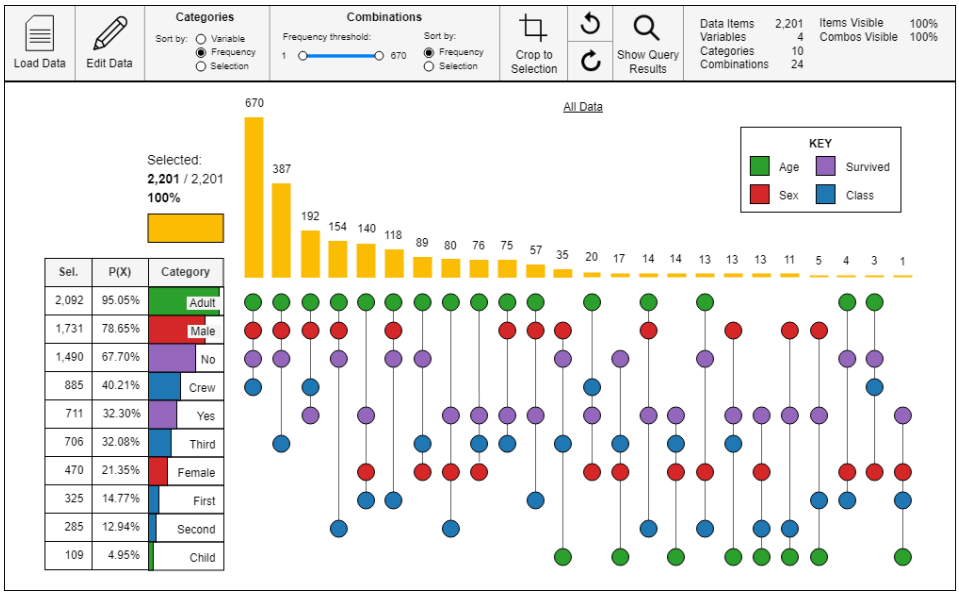


Figure 5.9: An early design of MultiCat, which looked less like a spreadsheet and more like a custom set-based representation (Trye, 2022).

Fourthly, space is now allocated on a per-variable rather than per-category basis. This makes the layout more efficient, especially if the dataset contains one or more high-cardinality variables. Of course, the stickers in the new design are also much wider than their node predecessors, so laying them all out side-by-side would necessarily consume a lot more space.

The fifth change concerns the assignment of colour to variables. Originally, each variable was assigned a distinct colour, which was then shared among all categories belonging to that variable. This approach made it easy to differentiate between variables, but difficult to distinguish the categories within each one. To address this, we experimented with using different shades of the same hue for categories belonging to the same variable. However, this conflicted with having different opacities for selected and non-selected items; for instance, it was difficult to distinguish lighter categories from non-selected ones.

Another drawback of using different shades of the same hue was that this led to unintended salience effects. Darker shades appeared more influential within combinations, despite all categories in a combination having the same frequency. They also stood out disproportionately in infrequent combinations. Furthermore, this method of colour allocation only works for categories with a small number of categories as it is difficult to distinguish more than three shades of the same colour (Muth, 2021).

To overcome these issues, we decided to allocate the same set of colours

to each variable, following the same order of assignment. This method aligns with the default behaviour of TableLens (Rao and Card, 1994) and Taggle (Furmanova et al., 2020). Since we sort the categories within each variable, this also implicitly conveys their relative rankings, albeit in a less intuitive way than a sequential scale.

Our sixth and final change was to add the column for residuals next to the frequencies. Inspired by UpSet's (Lex et al., 2014) "Deviation" measure, this shows the extent to which combinations are over- or under-represented within the dataset, which may lead to additional insights.

5.6 Interaction

This section describes the rich interactive features supported by MultiCat, which can be used in conjunction to highlight salient patterns, trends and relationships in the data. Users are initially presented with a high-level overview of categories and their combinations, but they may wish to interactively explore the data to gain a deeper understanding, either in a directed or undirected manner. The use of common spreadsheet operations, such as sorting and reordering columns, helps to consolidate the user's sense-making process (Bartram et al., 2021), while features such as selection and filtering enable rapid visualisation and comprehension of user-defined queries.

5.6.1 Sorting

MultiCat's sort functionality, which relates to design requirement R8, can be used to reveal relationships between category subsets and combinations. By default, combinations (rows) in the spreadsheet view are sorted by descending frequency, with residuals breaking ties (see Figure 5.5). Interacting with column headers rearranges the combinations, and through such exploration, enables the user to discover potentially revealing ways of viewing the data. Each column header is marked with a small triangle, indicating its ability to be sorted and reflecting the current state of the display. The triangle of the most recently sorted column is black, whereas all others are light grey. A single click on a column sorts it in descending order, with the sorting method varying by data type: nominal variables by rank frequency; ordinal variables by the number preceding the category name; and frequency and residual columns by numeric value. Clicking again on the same column switches to ascending order (Figure 5.7). For nominal and ordinal variables, this sorting mirrors or inversely matches the top-to-bottom order of categories in the sidebar. Sort-

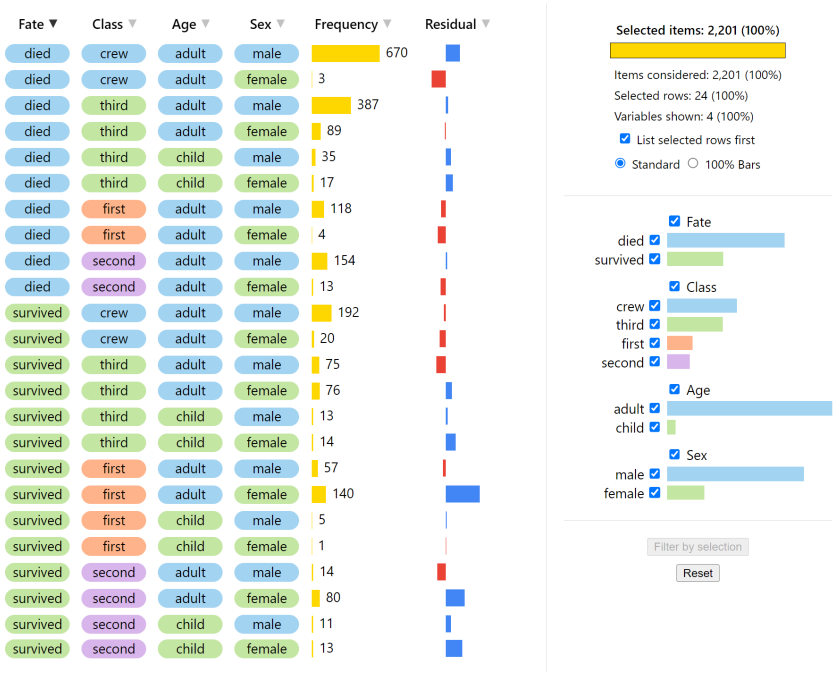


Figure 5.10: The Titanic dataset with combinations sorted by all four categorical variables. The columns have also been reordered, with Fate now appearing on the left.

ing by frequency and residual columns quickly highlights minima, maxima and outliers. Multi-column sorting is enabled by clicking on several column headers in succession, creating a user-defined hierarchy where the last sorted variables are prioritised. This is evident in Figure 5.10, where the four categorical variables have been sorted from right to left, with Fate at the top of the hierarchy. Note how the frequencies and residuals fluctuate considerably from row to row.

The order of combinations (rows) in the spreadsheet view can be configured to either prioritise the user's current selection or remain independent of it. The sidebar features a "List selected rows first" checkbox, which is enabled by default. When this option is selected, highlighted combinations are grouped at the top of the display, with the current sort criteria being applied separately to selected and non-selected items, as demonstrated in Figure 5.11. Conversely, when this option is unchecked, all rows follow a global sort order, regardless of selection status, as depicted in Figure 5.12. These settings emphasise different aspects of the data: the former facilitates direct comparisons of combinations of interest, while the latter reveals their distribution within the broader context of the dataset.

5.6.2 Reordering

Columns representing categorical variables can be reordered by dragging and dropping their headers to a new position. The top-to-bottom ordering of categorical variables in the sidebar matches their left-to-right ordering in the spreadsheet view. This means that variables further to the left in the spreadsheet view appear higher in the sidebar. Users may wish to organise variables in a specific manner, or ensure that the response variable occupies the left-most column, so that it is prominent in both views. The columns in Figure 5.10 have been rearranged so that the response variable (Fate) comes first rather than last.

5.6.3 Customising category bar charts

The two radio buttons in the sidebar control the appearance of the individual category bar charts. The default "Standard" option scales each bar's length according to the most frequent category in the dataset (e.g., "adult" in the Titanic dataset). In contrast, the "100% Bars" option normalises bar lengths, so that the selected proportions of each category can be directly compared, as shown in Figure 5.12. This feature—reminiscent of Mondrian's (Theus, 2002) built-in support for converting bar charts to Spine plots—effectively conveys part-whole relationships and is particularly useful for visualising relatively infrequent categories, which might otherwise be difficult to discern.

5.6.4 Brushing and linking

MultiCat uses brushing and linking (Hearst, 1999) to capture the association between selected items in the spreadsheet view and sidebar. Selected items in MultiCat are fully opaque, whereas non-selected items are rendered partially transparent to reduce their salience. This is exemplified in Figure 5.11, which highlights female adults in the Titanic dataset. Whenever a selection is made, three updates occur simultaneously: matching combinations (rows) in the spreadsheet view are highlighted; the statistics in the sidebar are updated, with the yellow chart showing the selected items as a proportion of the current filter; and the individual category bars in the sidebar reflect the corresponding proportion within each category. Together, these features play a critical role in revealing complex interactions between multiple categories and their combinations, helping users to see higher-dimensional features and structures in the data.

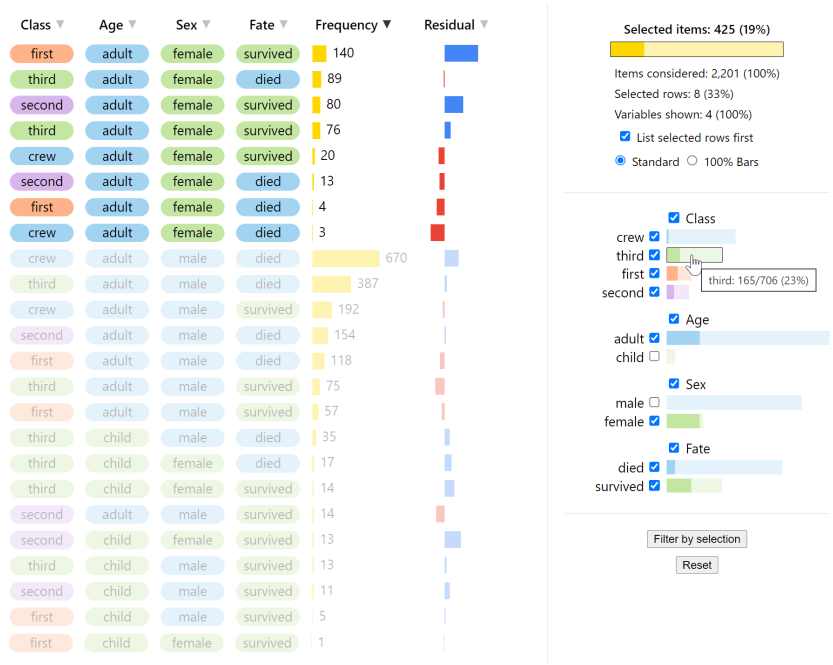


Figure 5.11: The Titanic dataset with all 425 female adults selected. The "Selected items" bar chart indicates that the joint probability of someone on the Titanic being female and an adult is 19%.

5.6.5 Tooltips

Hovering over most visual elements in MultiCat produces an informative tooltip, yielding "details-on-demand" (Shneiderman, 1996). In the spreadsheet view, hovering near a yellow frequency bar displays that combination's relative contribution to the current filter, whereas hovering near the red and blue residuals shows their exact values. Explanations of these two metrics are accessible via tooltips associated with their column headers. Additionally, tooltips reveal the full names of items on column headers or stickers, which is helpful for truncated text. In the sidebar, tooltips detail the selected proportion of each category in the format "third: 165/706 (23%)" (see Figure 5.11), where the numerator shows the exact number of selected instances and the denominator reflects the total number of occurrences of the category within the current filter. Finally, tooltips for the inactive "Filter by selection" button explain the criteria for its activation.

5.6.6 Dynamic queries

In MultiCat, dynamic queries facilitate the exploration of specific groups of categories, providing a logical and intuitive means for users to drill down into the data. These queries integrate interactive refinement and visual feed-

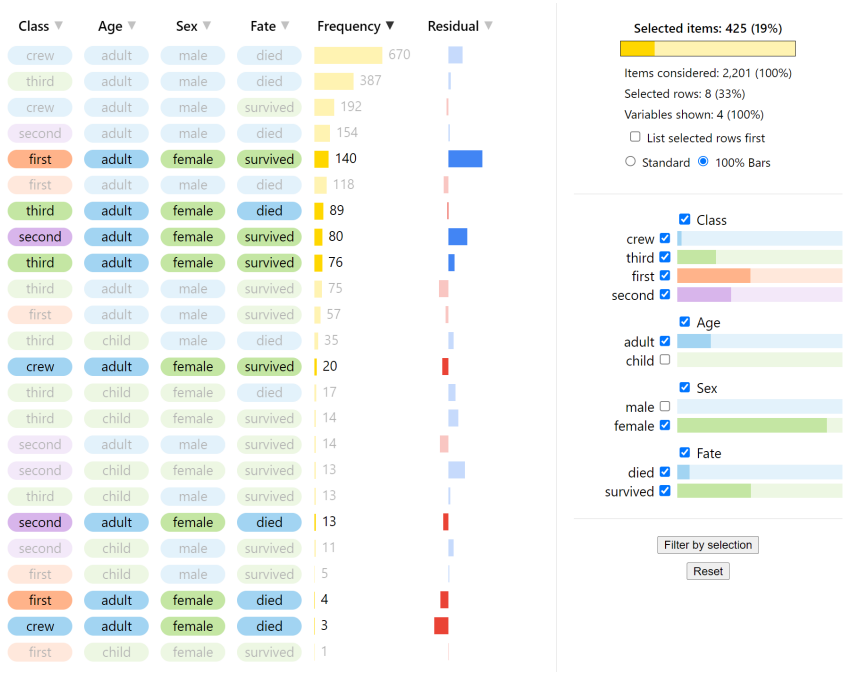


Figure 5.12: The Titanic dataset highlighting female adults, as per Figure 5.11, but with the "List selected rows first" checkbox disabled and 100% bar charts displayed in the sidebar.

back (Shneiderman, 1994), as per requirement R6. Users can form simple Boolean queries involving AND/OR logic by manipulating the category checkboxes in the sidebar. These checkboxes can be toggled directly, or the user can click on or besides the category bars to select *only* that category from within its parent variable. This is a useful shortcut for isolating one or a few categories within a high-cardinality variable. Alternatively, the user can click on category stickers in the spreadsheet view; this has the same effect as toggling the checkboxes, unless all categories for the parent variable are already selected, in which case it acts like the bar shortcut.

MultiCat employs straightforward Boolean logic in its queries: it uses OR (union) logic for categories within the same variable and AND (intersection) logic across different variables. This design prioritises simplicity over expressiveness, maximising ease of use and reducing the risk of logical errors (Spoerri, 1995). It leverages the principle that AND-ing categories within the same variable, under mutual exclusivity, always leads to an empty intersection (i.e., no matching records). Generally, selecting more categories in MultiCat broadens a query's scope, while choosing fewer categories narrows it. Each represented variable must have at least one selected category for matches to occur. Currently, it is not possible to formulate complex queries in MultiCat that feature multiple levels of nesting or incorporate more sophisticated Boolean operators

such as XOR.

Dynamic queries allow users to adjust their selection based on their information needs. The interactive and exploratory nature of these queries encourages users to ask questions of the data that they might not otherwise consider, such as "Are there more items with characteristics X than Y?" or "What happens if I select or deselect this checkbox?" In this process, users may uncover strongly associated categories, or one-way dependencies where a less frequent category is nearly always accompanied by a more frequent one, but not vice versa.

The "Selected items" bar chart functions as a real-time probability calculator for the current query, displaying empirical values based on actual data observations. This chart shows *a priori* probabilities when no filter is applied ("Items considered" is 100%) and conditional probabilities otherwise. For instance, Figure 5.11, in which all data items are represented, shows the proportion of women (adult + female) on board the Titanic.

As seen in Figure 5.1, Mondrian allows users to formulate Boolean queries through linked bar charts (Theus, 2002), similar to MultiCat. While Mondrian offers a wider range of Boolean operators than MultiCat, and allows more flexibility in applying them—for instance, OR-ing categories from different variables is permitted—it lacks any visual clues as to a query's internal representation. This can easily lead to user errors, especially when constructing complex queries involving multiple variables and operations. In contrast, MultiCat's simple design ensures that the syntax of a query can always be inferred from the active checkboxes in the display.

5.6.7 Filtering

Filtering is a useful strategy for reducing the size and complexity of a categorical dataset. In MultiCat, it is possible to filter out data items that are not part of the current category selection, as well as entire variables. This aligns with design requirements R6 and R7. Filtering differs from selection in that excluded items are removed from the display, rather than merely being faded out.

Category-based filtering is accomplished by selecting a subset of categories and clicking on the "Filter by selection" button; this removes non-selected combinations (rows) from the spreadsheet view and the corresponding items from the sidebar. To prevent confusion, categories always remain the same colour, even if their ranking changes within a new filter. However, the order of categories within the sidebar updates accordingly. Filtering criteria can be

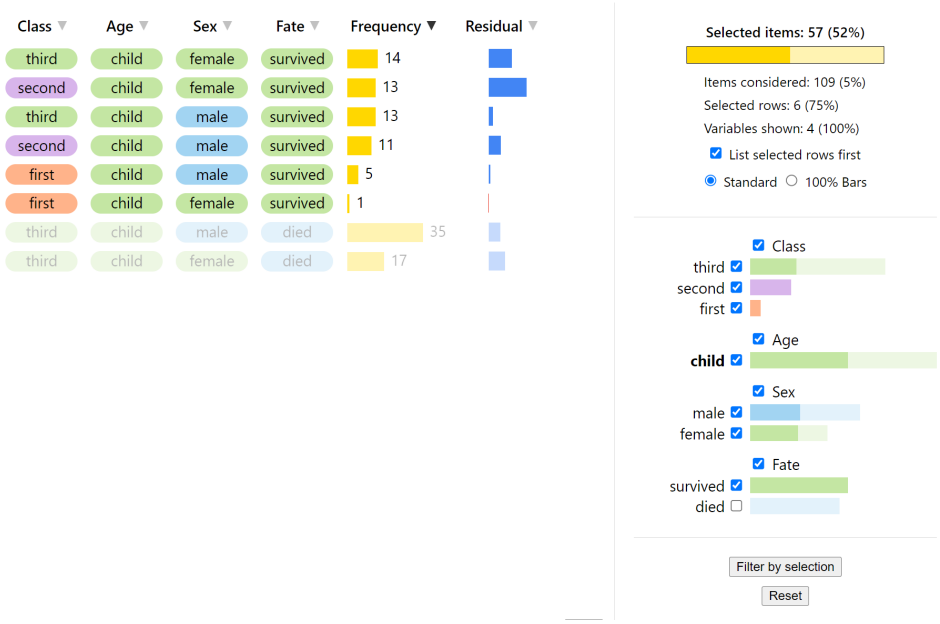


Figure 5.13: The Titanic dataset filtered by children and showing the conditional probability $P(survived|\mathbf{child})$.

progressively refined to reveal more in-depth relations in the data. Categories that are no longer represented in the filtered data are removed from the sidebar to save space, whereas categories that form the basis of the filter are emphasised in bold; these correspond to the "given" part of a conditional query.

As an example, Figure 5.13 shows the Titanic dataset filtered by "child". Since children accounted for only a small proportion of passengers, their distribution is quite different from the overall dataset shown in Figure 5.5. For instance, there are many more children in second class than first class, and there is a relatively even split between children who died and survived. The query in Figure 5.13 specifically shows the conditional probability $P(survived|\mathbf{child})$; that is, the percentage of children who survived: 52%, or 57 children as an absolute value. The opacity of the bars shows that all children in first and second class survived (although there were relatively few children in these classes), whereas only a third of children in third class survived. Notably, while a similar *number* of boys and girls survived, a greater *proportion* of boys died, which is also true (and more pronounced) for males and females in general.

In addition to category filtering, MultiCat supports the removal of variables via the variable checkboxes in the sidebar, which are positioned centrally. The user can examine as many or as few variables at a time as desired. When variables are removed, combinations in the spreadsheet view are instantly updated, including their frequencies and residuals. The category bars for disabled variables are removed from the sidebar, but the variable checkboxes remain visible

so that they can be re-selected. Generally, reducing the number of variables increases combination frequencies in the spreadsheet view as there are fewer permutations; however, the number and distribution of categories within these variables is still a significant factor.

5.6.8 Scrolling

Scrolling becomes necessary when there is too much data to fit everything in the available space. If there are too many combinations (rows) in the spreadsheet view, a vertical scrollbar appears in the main window. The header row of the table is frozen, so that the column names remain visible when the user scrolls. Likewise, a horizontal scrollbar is added when there are too many variables (columns). As previously mentioned, an internal vertical scrollbar is added to the sidebar when there is insufficient space to show all category bars at once. The summary statistics and buttons in the sidebar are also fixed in place. Datasets that require excessive scrolling in either direction increase the user's cognitive load and are therefore not recommended, though the user may be able to quickly derive meaningful subsets using the above interactive features. Zooming out may also help in some cases.

5.7 Implementation

A live demo of MultiCat, preconfigured with the Titanic dataset (Dawson, 1995) for demonstration purposes, is available at `https://dgt12.github.io/multicat/`. Researchers wishing to visualise their own categorical datasets can do so by downloading the source code, replacing the input file, and running the application on a local server. Detailed instructions and all necessary code are provided in the project GitHub repository: `https://github.com/dgt12/multicat`.

MultiCat is implemented in Svelte (`https://svelte.dev/`) and has been tested in Google Chrome with datasets containing up to 20 categorical variables and 500,000 data items. The data are currently read in as an array of JavaScript objects, where each object represents a single item. All data are converted from case form to frequency form before the visualisation is rendered. The spreadsheet view is simply an HTML table.

Regarding limitations, our prototype does not currently integrate with other tools, support more than one input format, provide edit history (undo/redo) or generate publication-ready figures. There are also important scalability constraints, which we discuss below.

5.8 Scalability

MultiCat is built on the premise that aggregation is crucial for creating scalable visualisations. Consequently, its efficacy depends on the presence of many recurrent combinations of categories; this is what exposes meaningful structure within the spreadsheet view. As the number and diversity of categories increase, combination frequencies typically decrease, except in cases where variables are highly associated. Introducing a single high-cardinality variable with weak associations drastically reduces the number of recurrent combinations in MultiCat, underscoring the importance of strategic variable selection. Nevertheless, even in datasets comprising predominantly unique records, MultiCat's interactive features, coupled with the sidebar, remain valuable for exploring the data in a task-driven manner.

The screen space needed for a MultiCat visualisation is a function of the number of categorical variables (horizontal space) and the distinct combinations they form (vertical space). As with many other techniques for visualising categorical data (Hofmann, 2006), MultiCat's display is relatively independent of the number of data items. The maximum number of combinations (rows) for a given set of variables can be calculated by multiplying the variables' cardinalities. This represents the worst-case scenario for vertical space allocation, but it is unlikely that every possible combination will appear in a real-world dataset with several variables. While scrolling allows every combination to be viewed if necessary, other features such as sorting, filtering and querying allow the user to better utilise the available screen space.

Theoretically, design aspects such as scrollbars, fixed column widths and systematic allocation of colours mean that MultiCat can accommodate any number of categories and variables. In practice, however, having too many of these can lead to highly fragmented groups in the spreadsheet view, limiting its effectiveness. Moreover, MultiCat visualisations become challenging to interpret when the number of variables exceeds screen capacity. While scrolling allows access to additional variables, perceiving differences among category combinations becomes difficult as users must remember parts of the display that are not immediately visible.

In terms of cardinality, for best results, most variables should have only a handful of categories, with none exceeding ten. MultiCat is capable of handling datasets that go beyond these limits, as per design requirement R4. However, operating within the stated limits not only aids in keeping recurrent combinations together, but also ensures the number of distinct grey stickers per variable is minimised. This is helpful since the grey stickers are harder

to differentiate than those with different colours. Additionally, this approach declutters the sidebar, reducing the need for extensive scrolling and enabling easier comparison of variable distributions.

5.9 Formative user study

This section outlines the procedure and findings of our small-scale user study of MultiCat.[1] The study aimed to detect general usability issues, collect participant feedback, improve the prototype and assess users' ability to interpret the visualisation without prior training. Based on the outcomes of the study, we made several refinements to the prototype, which are detailed below.

5.9.1 Procedure

Our study employed a within-subjects design. Six participants with a background in computer science were recruited, including four students and two staff members. Following other research, this was deemed to be enough participants to identify significant usability issues in a cost-effective manner (Peña-Araya et al., 2022). All participants completed the study in the University of Waikato's Usability Lab using the same Windows 11 laptop and mouse. The study was divided into an *Exploratory Phase*, where participants familiarised themselves with the layout and functionality of the MultiCat prototype, and a *Task Phase* comprising the same seven tasks for each of two datasets. Subsequently, participants completed a short online questionnaire. They were encouraged to think aloud throughout the study, and a screen and audio recording was captured for detailed analysis. Each session took approximately 30 minutes, with participants giving informed consent at the beginning.

During the initial *Exploratory Phase*, participants were introduced to the MultiCat prototype with the Titanic dataset loaded in (Dawson, 1995). Having only been told that MultiCat was designed for visualising multiple categorical variables, and that they were viewing data related to the Titanic disaster, participants were asked to explain what they thought the visualisation was showing. They were then encouraged to interact freely with the prototype. The interviewer facilitated this exploration by answering questions and guiding participants towards features they had not yet encountered, using prompts such as "What happens when you hover over one of these frequency bars?" This approach ensured that participants actively engaged with and verbalised their understanding of the functionality, rather than simply being told what

[1] Ethics approval for this study is given in Appendix B.

it did. This in turn provided valuable insights into which features were and were not intuitive.

In the subsequent *Task Phase*, participants were asked questions about two publicly available datasets of varying complexity. First, they revisited the Titanic dataset from the *Exploratory Phase*, which comprises four variables, 10 categories and 2,201 observations. The second dataset was a simplified version of the Mushroom dataset (Schlimmer, 1987), representing a hypothetical collection of 8,124 mushrooms belonging to 23 species. We selected only eight of the original 22 variables to ensure all columns were visible in MultiCat without the need for scrolling (see Figure 5.14). These variables, featuring between two and seven categories for a total of 34 categories, encompass a diverse range of properties, including the mushrooms' edibility, physical characteristics, population and habitat. As neither dataset in the study incorporates ordinal variables, we did not evaluate MultiCat's capabilities for handling such data.

Table 5.1 details the tasks that participants were asked to carry out, including the specific questions posed for each dataset. These tasks varied in complexity and were chosen to reflect common activities in categorical data analysis. The first task, while straightforward, served to acquaint users with the size and structure of the dataset. Other tasks integrated different visual elements and features, enabling observation of user strategies. Participants did not receive any feedback on their answers and were asked to reset the display between tasks to ensure a fresh start each time. To mitigate potential memorisation effects, the sequence of tasks for the Mushroom dataset was pseudo-randomised.

After completing the *Task Phase*, participants filled out an online questionnaire. The first set of questions in this survey was about users' familiarity with related tools and their knowledge of statistical concepts, while the next set required them to provide a subjective rating of their experience and impressions using MultiCat. Finally, there were three open-ended questions asking users to identify things they liked and disliked about MultiCat, as well as any suggestions they had for enhancing the interface.

Table 5.1: Task descriptions and associated questions for the Titanic and Mushroom datasets.

Task	**Titanic Dataset (Dawson, 1995)**	**Mushroom Dataset (Schlimmer, 1987)**
T0: Summarise dataset	How many items (in this case, *people*) does the dataset contain? How many categorical variables does the dataset contain?	How many items (in this case, *mushrooms*) does the dataset contain? How many categorical variables does the dataset contain?
T1: Identify key N-way relationship(s)	What is the most frequent combination of categories involving all variables and how often does it occur? What proportion of the total dataset does this combination account for?	How often do the most frequent combinations of categories involving all variables occur? How many combinations with this frequency are there? Do they share any of the same characteristics? If so, what are they?
T2: Find absolute value and (marginal) frequency for a category or subset of categories	How many children were on board the Titanic? What percentage of the data do the children account for?	How many mushrooms have a pendant ring type? What percentage of the data do they account for?
T3: Compare frequencies of categories or subsets involving different variables	Which category is *more* frequent: "female" or "first" class?	Which category is the *least* frequent out of "convex" cap-shape, "broad" gill-size and "no" bruises?
T4: Find non-conditional probability	What proportion of people on board the Titanic were female passengers (i.e., non-crew) who survived?	What proportion of mushrooms are edible, have a convex or flat cap, and reside in scattered populations?
T5: Find conditional probability	What is the probability (as a percentage) that someone was in first class, given that they were female?	What is the probability (as a percentage) that a mushroom does not have a smooth stalk surface, given that it is edible and has no bruises?
T6: Explore a (binary) response variable with respect to all other variables	Let's say you are particularly interested in the people who survived the Titanic disaster. Do you notice any trends among this group of people? How about with respect to over- or under-represented groups?	Assume you are particularly interested in edible mushrooms, and you want to avoid the poisonous ones. How many edible mushrooms are there? For which categories/properties can you be certain that a mushroom will be edible rather than poisonous?

Figure 5.14: The Mushroom dataset highlighting items that are edible, have a convex or flat cap and reside in scattered populations, as per task T4.

5.9.2 Results

We now detail the results of our user study, providing general observations from each phase, a discussion of factors influencing task completion, and a summary of participants' responses to the post-study questionnaire.

The self-reported prior knowledge of our six participants is presented in Table 5.2. All participants indicated at least moderate familiarity with categorical data, with half of them reporting high familiarity. They all described themselves as being very familiar with spreadsheet applications, such as Microsoft Excel and Google Sheets, and possessed at least a basic understanding of statistical concepts. Finally, while most participants considered themselves moderately familiar with visualisations, none identified as an expert in this area.

Table 5.2: Summary of participants' self-reported familiarity with relevant tools and concepts.

Topic (1=unfamiliar, 5=extremely familiar)	**Median**	**Mode**
Bar charts	3.5	4
Spreadsheet applications (Microsoft Excel, Google Sheets)	4	4
Visualisations (in general)	3	3
Categorical data	3.5	3, 4
Joint, conditional and marginal probabilities	3	3
Observed frequencies, expected frequencies and deviations	3	3

In the *Exploratory Phase* of the study, five of the six participants quickly and accurately described the key features of MultiCat by themselves, while the sixth participant needed some guidance. For example, one participant commented within a matter of seconds "Ahh, so this [row] is like the combination, so it's saying 670 people were crew, adult, male and died". Participants also made relevant observations about the sidebar: "I see you've colour-coded the values and this [sidebar] is like a legend to go with it" and, after making a selection, "I would imagine this filled in bit [of each bar] is the data that's being actually used, and the whole thing is the total amount of data". Participants explored several interactive features on their own, often correctly deducing that it was possible to sort the data by clicking on the column headers and that deselecting the category checkboxes in the sidebar would remove them from the selection.

Regarding points of confusion, a few participants tried sorting the spreadsheet view by multiple columns but did not find this process intuitive. For example, one participant described the sorting order as "back-to-front". Two participants tried clicking on the category stickers, expecting this to filter the data, but this feature had not yet been implemented. There were a few instances of "change blindness", whereby users made a selection and immediately noticed that the combinations in the spreadsheet view had changed, but not the content in the sidebar. However, once they realised the two views were linked, this greatly enhanced their understanding of the interface. The "Filter by selection" button was another source of confusion, as it had no effect when participants clicked it without having made a selection. As one participant noted, "I find the filtering a little confusing, but I think if I used it and played with it, it would make more sense". Most participants incorrectly assumed that the "Deviation" metric—which is what the "Residual" column was previously called—was based on the standard deviation, until its actual function was clarified. Finally, the tooltips were quite delayed, which resulted in some participants missing relevant information on their first attempt to hover over different components.

Figure 5.15 summarises results from the *Task Phase*, broken down by dataset and participant (P1-P6). For the most part, participants were able to complete tasks quickly and successfully, but they encountered similar issues and consistently struggled with tasks T5 and T6. Among the 14 task iterations, the number of correct answers per participant ranged from 9 to 13, with everyone succeeding at tasks T0, T2 and T4 across both datasets. For correctly solved tasks, participants mostly used the expected strategies given

in Appendix C. During the first iteration of T6, for example, participants extracted trends relating to Titanic survivors by first selecting the 'Survived' category and then detecting patterns in the spreadsheet view, sometimes sorting by categories and/or residuals to facilitate this process.

Task	Dataset	P1	P2	P3	P4	P5	P6
T0	Titanic						
	Mushrooms						
T1	Titanic						
	Mushrooms						
T2	Titanic						
	Mushrooms						
T3	Titanic						
	Mushrooms						
T4	Titanic						
	Mushrooms						
T5	Titanic						
	Mushrooms						
T6	Titanic						
	Mushrooms						

Figure 5.15: Matrix of user study results showing tasks as rows, differentiated by dataset, and participants as columns. Blue cells signify correct responses, yellow cells denote partially correct responses (right approach, wrong answer) and red cells signify incorrect responses.

When selecting categories, five participants effectively used the bar shortcut in the sidebar, while the remaining participant preferred to toggle the checkboxes individually. The use of sorting varied among participants, with some relying on it quite heavily and others not using it at all. One participant incorrectly answered two questions in the Titanic dataset after inadvertently scrolling past the top two combinations. At that time, the "Reset" button did not reposition the scrollbar, which meant this had a flow-on effect (this issue has since been addressed; see *Refinements* below).

Participants were sometimes uncertain which part of the interface they should use for specific tasks. They tended to focus on the spreadsheet view, even for T2 and T3, which involved univariate category frequencies and were therefore better suited to the sidebar. This was also the case for T6 in the Mushroom dataset, where participants needed to identify categories unique to

edible mushrooms. Most participants attempted to answer this question by manually sifting through the combinations to find categories that were present in the "edible" selection, but absent from the non-selected (i.e., poisonous) data. While entirely possible, a much faster strategy—which one participant employed—was to look for fully opaque categories in the sidebar after selecting "edible" mushrooms.

Another observation is that participants sometimes hid variables that were not directly related to a task's requirements, especially within the Titanic dataset. This may have been motivated by a desire to simplify the visualisation as much as possible. For example, in task T5, which asked about the proportion of females in first class, one participant excluded the variables Age and Fate. While not incorrect, this was not necessary for completing the task, as retaining all categories for non-mentioned variables would not affect the relevant details in the sidebar. In practical scenarios, removing variables can be counterproductive as it reduces the dimensionality of combinations in the spreadsheet view, obscuring potential insight into more complex relationships. However, this did not matter within the context of our study, especially since the display was reset after each task.

The task with the lowest success rate was T5, which required participants to calculate a conditional probability. The most common approach was to select the mentioned categories without applying a filter, then read the resultant probability from the "Selected items" bar. This yielded the correct numerator but an incorrect denominator, leading to an incorrect answer. The "Filter by selection" feature was largely overlooked, being used by only two participants. This perhaps reflects a gap in participants' understanding of conditional probabilities, although there is clearly also room for supporting these better. Interestingly, one participant with a strong background in statistics extracted the numerator and denominator for each conditional probability from the non-filtered display, choosing to give their answer as a fraction. The same participant sometimes manually added the frequencies of relevant combinations rather than selecting them in the sidebar.

Overall, participants performed better with the Titanic dataset and found it much easier to navigate than the Mushroom one. This was to be expected, given that the Mushroom dataset had significantly more categories and variables, and required vertical scrolling in both views. Participants' familiarity with the category-variable relationships in each dataset may have been another important factor, though our study did not control for this.

In the questionnaire, participants rated MultiCat very highly, as shown by their responses in Table 5.3. Most participants commented in the open-ended questions that they liked the appearance of MultiCat and found the interface easy to use and understand. For instance, one participant remarked "MultiCat is very intuitive. I really enjoyed the visual aspects, being able to visually see the categories, the relationships between them, etc.", while another stated "The visualisations made it easy to see data at a glance. The tooltips were really helpful."

Regarding things they disliked, three participants noted that they found it comparatively difficult to navigate the Mushroom dataset, with one participant saying "I guess I found it a little hard to answer questions with the mushroom dataset in terms of finding the categories". However, as another participant observed, this is to be expected when analysing more complicated data: "The more complex interactions were a little tricky on the first try. BUT this is allowing you to visualize and analyze more complex relationships, so it makes sense that it wouldn't be as straightforward as the more simple visualizations. I could imagine this being an incredibly useful tool!"

There were three suggestions for improving MultiCat: (1) provide more informative tooltips; (2) allow the user to formulate queries using the category stickers in the main visualisation; and (3) allow dynamic resizing of column widths to view the full variable and category names, without having to inspect the tooltip. These first two suggestions have been incorporated into the updated prototype, as noted in *Refinements* below.

Table 5.3: Summary of participants' responses to different statements about MultiCat. Values marked with an asterisk have been adjusted to enable direct comparison with other questions, where higher values are better.

Statement (1=strongly disagree, 5=strongly agree)	**Median**	**Mode**
I found MultiCat easy to use.	4	4
I was able to complete the tasks.	4	4
I felt confident using MultiCat.	4	4
I thought that the main visualisation and the sidebar worked well together.	5	5
I thought some features were unnecessarily complicated *(lower is better)*.	2 (4*)	2 (4*)
I thought the interactive features (sorting, querying, filtering) were useful.	5	5
I thought the interactive features worked well together.	4.5	4, 5
I think that I would need assistance to use MultiCat again *(lower is better)*.	2 (4*)	2 (4*)
I think most people would learn to use MultiCat fairly quickly.	4	4
I would like to use MultiCat again in the future.	4.5	4, 5
Overall rating (1=unusable, 5=exceptional)	5	5

Overall, reflecting on our study, participants found the concept of MultiCat compelling and were enthusiastic about using it again in the future. They succeeded in performing a wide range of tasks, but clearly found some features (like filtering) less intuitive than others. While some issues with the prototype were identified, these do not overshadow MultiCat's potential as a valuable tool for analysing categorical data.

5.9.3 Refinements

Based on observations and feedback elicited from the user study, we made the following changes to the MultiCat prototype, which were already accommodated in our prior explanation of the technique:

1. Renamed the "Deviation" column to "Residual" to avoid confusion with the standard deviation.
2. Added more informative tooltips to the "Frequency" and "Residual" column headers.
3. Extended the query functionality to allow clicking on category stickers within the spreadsheet view.
4. Adjusted the scaling of the category bar charts in the sidebar to enable direct comparison across different variables. Previously, the bars were scaled according to the most frequent category within each variable, rather than the global maximum.
5. Added the two radio buttons to the sidebar, instead of just offering the "Standard" view for category bar charts.
6. Greyed out the "Filter by selection" button when it has no effect.
7. Modified the "Reset" button to reconfigure the vertical scrollbars for the spreadsheet view and sidebar.

5.10 Comparison with existing techniques

In this section, we compare the strengths and weaknesses of MultiCat with two existing techniques: Parallel Sets (Kosara et al., 2006) and (Interactive) Mosaic Plots (Hartigan and Kleiner, 1981; Theus, 2002). We have chosen these techniques for three reasons: (1) they are established methods for visualising multidimensional categorical data; (2) they directly encode cells in contingency tables, rather than employing dimensionality reduction techniques, meaning they are CatViz, not QuantViz, techniques; (Johansson Fernstad and Johansson, 2011) and (3) they preserve higher-order relationships, unlike, for instance, the Heatmap Matrix (Rocha and da Silva, 2018; Trye et al., 2023), Mosaic Ma-

trix (Friendly, 1999), or GPLOM (Im et al., 2013), which only explicitly show pairwise relationships. While MultiCat meets these last two criteria, it differs from Mosaic Plots and Parallel Sets in that it does not use an inherently hierarchical or area-proportional layout. This has important implications for its relative strengths and weaknesses, as discussed below.

All three techniques—MultiCat, Parallel Sets and Mosaic Plots—facilitate quick identification of the most frequent combinations of categories involving N variables. In MultiCat, these combinations are prominently displayed in the top rows of the default spreadsheet view; in Mosaic Plots and Parallel Sets they are shown by the largest tiles and largest parallelograms in the bottom "tier", respectively. Of these techniques, Mosaic Plots have the most intuitive semantic structure, as combinations involving subsets of categories are logically laid out side-by-side. Mosaic Plots are also unique in that the tiles align when variables are independent; (Friendly, 1999) this cannot be so easily discerned from MultiCat or Parallel Sets. However, at the same time, Mosaic Plots scale poorly when there are more than four variables because this means more than two variables have to be plotted on the same axis, increasing the potential for confusion.

MultiCat excels at helping users to identify outliers, namely combinations that were only observed once or a handful of times. By default, these combinations are situated at the bottom of the spreadsheet view, but they can be easily brought to the top by reversing the sort applied to the "Frequency" column. In contrast, identifying such rare combinations in Mosaic Plots and Parallel Sets is more challenging due to their area-proportional layouts, which result in very small tiles and parallelograms. MultiCat overcomes this limitation with its tabular layout, where all rows are of uniform height, guaranteeing their readability.

One limitation that MultiCat shares with Parallel Sets is the inability to display non-observed combinations (i.e., those with a frequency of 0). In certain situations, including sanity checks, it is useful to identify or estimate the number of non-occurring combinations. Some implementations of Mosaic Plots address this by representing non-occurring combinations with a small circle, making them distinguishable (Hofmann, 2000).

As alluded to above, both Parallel Sets and Mosaic Plots employ a hierarchical layout, but they do so in distinct ways. These hierarchies emphasise conditional relationships between variables. In Parallel Sets, the order in which the subsets are derived can easily be ascertained by following the variables from the top tier down to the bottom. Only the bottom tier of a Parallel Sets

visualisation shows relationships involving all variables simultaneously. The upper tiers can be useful for revealing interactions among fewer variables, but they occupy additional space and privilege variables that are higher up, potentially biasing the viewer's interpretation. Furthermore, changing the order of variables and/or categories alters the appearance of the display, sometimes drastically, which can in turn influence the insights derived. Similarly, with Mosaic Plots, the order in which variables are split affects what can be seen in the visualisation (Hofmann, 2006). The order in this case is less obvious than in Parallel Sets, but can be deduced from the category labels usually found along the external edges of the display. However, tools like Mondrian only display category labels for the (two) outermost variables, necessitating the use of interactive tooltips to identify labels for nested variables. This can make it difficult to perceive the full structure of the nested data at a glance (Hofmann, 2000).

In contrast, MultiCat allows users to discern patterns without being constrained by a predefined hierarchy. Although the order of categorical variables (columns) might subtly influence the interpretation of combinations, it does not change the content of each combination and therefore does not profoundly impact the display. The initial order of combinations (rows) is determined by the combination frequency rather than by related groups of categories. This approach ensures that variables are treated as equally as possible, unless the user decides to sort the combinations by one or more categorical variables, thereby specifying a hierarchy of their own. If users do want to focus on a particular subset of variables, they can query the data or filter out certain variables. MultiCat thus promotes user-driven exploration of important variables and relationships in a relatively undirected manner.

Extracting complete combinations of categories is arguably more straightforward in MultiCat than either Parallel Sets or Mosaic Plots. MultiCat's use of coloured stickers explicitly names each category within a combination, and this feature remains effective even for large numbers of variables. In contrast, Parallel Sets and Mosaic Plots typically require interactive tooltips for decoding combinations as other strategies are cognitively demanding. Moreover, since tooltips can usually only be accessed one at a time, MultiCat is more efficient for comparing multiple combinations involving a large number of variables at the same time.

In general, it seems easier to accurately compare combination frequencies in MultiCat than the other techniques. The lengths of bars in MultiCat's frequency bar chart are easier to compare than the areas of differently sized

and shaped tiles in Mosaic Plots, or the varied angles of parallelograms in Parallel Sets. Parallel Sets also invariably suffer from line crossings, which create visual interference, particularly in datasets with a high diversity of categories and variables. Mosaic Plots, while free from line crossings, are hard to read when there is an abundance of small tiles. MultiCat circumvents these issues since additional combinations can be scrolled vertically if they do not fit in the available screen space, and additional variables can be scrolled horizontally.

Both MultiCat and Mosaic Plots incorporate Pearson residuals, which are valuable for determining whether particular combinations of categories are over- or under-represented in the data. In Mosaic Plots, these residuals are typically shown by applying discrete (Friendly, 1999) or continuous (Zeileis et al., 2007) colour shading to the tiles. This works well for large tiles but not for small ones as it is difficult to make out the colours. Moreover, the different use of size and colour may lead to misinterpretations of the data; for instance, if two tiles have the same colour but are drastically different sizes, a viewer may mistakenly believe the larger one has a larger residual. MultiCat achieves a more precise encoding for the residuals by using a diverging bar chart that is separate from the bar chart for frequencies. Any residuals that are difficult to see have smaller absolute values and are therefore less important.

To summarise, MultiCat is useful for identifying combinations of any frequency and Pearson residuals of any size. It differs from the other techniques because it uses a tabular, non-hierarchical layout that does not encode frequencies in an area-proportional way. Parallel Sets clearly shows the order in which a hierarchy is formed, but this may impact what the user perceives in the visualisation. It does not support Pearson residuals and is less efficient for identifying infrequent combinations. Mosaic Pots, on the other hand, do support Pearson Residuals, and exhibit unique features such as the alignment of tiles for independent variables. However, they present readability challenges for small tiles and their colour-based representation of residuals is less perceptually accurate than MultiCat's length-based encoding. Moreover, Mosaic Plots do not scale well to more than four categorical variables. Thus, while each technique has its merits, MultiCat's approach can be seen to offer a more versatile and user-friendly solution for exploring categorical data.

5.11 Possible extensions

We propose the following extensions to MultiCat, which we believe would further enhance its usability:

Direct data manipulation: MultiCat could provide a means of accessing and editing the raw data. Users should be able to modify or delete specific items, merge existing categories, derive new variables from existing ones, and so on.

Heterogeneous datasets: In practice, datasets often comprise both categorical and continuous variables, rather than being limited to only one type (Zhang et al., 2014). Recognising this, MultiCat could be extended to also handle continuous variables. One possible approach is to categorise all continuous variables into bins (Wickham and Hofmann, 2011; Rocha and da Silva, 2018), treating these bins as ordinal categories. This would allow them to be integrated into each combination of categories. Alternatively, drawing inspiration from tools like UpSet (Lex et al., 2014), continuous variables could be kept in their original form and presented alongside categorical combinations as aggregated visualisations, such as box-and-whisker plots. This would afford insights into the extent to which data items within and across combinations of categories vary with respect to their continuous characteristics.

Drill down to individual records: Akin to Taggle (Furmanova et al., 2020), a drill-down feature would enable users to explore individual data items—in mini scrollable tables, for instance—within the spreadsheet view. An expand/collapse icon could be positioned next to each category combination. Clicking on this icon would reveal unique identifiers about the corresponding records, such as passengers' names in the Titanic dataset (one row per record). Additionally, a text search feature could visually highlight matching items. A global "Expand/Collapse All" option could also be included, with the parent combinations always remaining visible.

Response variables: When analysing categorical datasets, users may wish to examine a response variable in relation to all other variables (Agresti, 2012). To facilitate this, the sidebar could incorporate a drop-down menu for selecting a response variable. This menu would list the names of all columns, with "None" selected by default. Upon choosing a response variable, the corresponding column would be removed from the spreadsheet view. Instead, each of its categories would be given individual frequency/proportion and residual columns appearing alongside each combination. To visually distinguish them from other variables, these new bars for the response variable could employ hatching patterns instead of colour. This feature would enhance users' under-

standing of how categories within the response variable are distributed among combinations of all other variables. For instance, in the Titanic dataset, selecting Fate as the response variable would help to answer questions about mortality rates, such as "Which combinations had significantly more fatalities than survivors (or vice versa)?"

More powerful queries: The MultiCat interface could be adapted to support more expressive queries, by taking inspiration from tools such as ComBiNet (Pister et al., 2023), 2dSearch (Russell-Rose and Gooch, 2018) and AI-STARS (Anick et al., 1989). Currently, users are not able to OR categories across different variables, or create compound queries of arbitrary complexity. Following ComBiNet, it might be helpful to offer a synchronised text-based representation of queries, allowing users to edit *either* the selected visual elements or corresponding text. Additionally, MultiCat could incorporate the ability to save and reload queries, and even allow logical operations to be applied to the queries themselves.

Automatic feature selection: Upon loading a dataset in MultiCat, a "Configuration Selector" tool could be introduced to assist users in strategically choosing variables to include in their analysis. This tool would offer a visual summary of key characteristics of different subsets of variables, providing insights and automatic recommendations about different possible analysis paths. For example, it could display ranked information about the number of distinct combinations for each set of variables and their median frequency. This would be particularly useful for identifying influential relationships, especially in cases where certain variables add undue complexity by fragmenting frequent category combinations. As discussed in the section on Scalability, the impact of a single variable can be substantial, particularly if it encompasses a large number of categories or exhibits significant variation with respect to other variables. Therefore, the "Configuration Selector" would be valuable not only for setting up the initial display, but also for guiding users in making informed decisions about which variables and categories to include in their subsequent explorations.

Missing values: Settings could be added to MultiCat to provide an overview of missing values across all variables and to filter these out in a controlled way.

5.12 Conclusion

This paper has introduced MultiCat, a novel visualisation technique for exploring multidimensional categorical data. MultiCat combines the strengths of a tabular layout with multiple coordinated views, supporting the user in rapid data observation, hypothesis testing and exploratory information seeking. MultiCat distinguishes itself from other techniques through its: (1) high readability of category labels, which notably includes high-dimensional relationships and low-frequency combinations; (2) non-hierarchical default layout; (3) visual summary of individual category contributions; and (4) separate treatment of nominal and ordinal variables. The spreadsheet view provides a comprehensive overview of multidimensional relationships, complemented by sorting operations that enable task-driven analysis of typical observations and outliers alike. The sidebar helps to bridge the gap between individual categories and multidimensional combinations by summarising category distributions and indicating proportions of selected subsets. Furthermore, dynamic queries in MultiCat enable fast computation of absolute values and empirical probabilities, providing a natural and intuitive means of drilling down into the data. We validated MultiCat by conducting a small-scale user study, in which participants rated their experience highly and successfully performed a diverse range of tasks. The results of this study suggest that MultiCat would be a valuable tool for data analysts, while hinting at its advantages over traditional techniques.

Future work could focus on implementing and evaluating the proposed extensions, from direct data manipulation to special treatment of response variables. An in-depth comparative study between MultiCat and established techniques for visualising multidimensional categorical data would also be valuable, in order to better understand the relative strengths and weaknesses of each approach for various analysis tasks.

Acknowledgements

We would like to thank our study participants for generously giving up their time to contribute to this research. We are grateful to Andreea Calude for her comments on the draft. DT thanks the University of Waikato for funding this research through a Doctoral Scholarship.

5.13 Postscript

In this chapter, we designed, implemented and evaluated MultiCat, a tabular technique for visualising multiple categorical variables simultaneously. We have explained some clear advantages of MultiCat over existing alternatives, which we will leverage in Part III (Chapter 8).

5.14 References

(2009). parsets. `https://code.google.com/archive/p/parsets/downloads`. Accessed January 25, 2024.

Agresti, A. (2012). *Categorical data analysis*. John Wiley & Sons, 3rd edition.

Alsallakh, B., Aigner, W., Miksch, S., and Gröller, M. E. (2012). Reinventing the Contingency Wheel: Scalable visual analytics of large categorical data. *IEEE Trans Vis Comput Graph*, 18(12):2849–2858.

Anick, P. G., Brennan, J. D., Flynn, R. A., Hanssen, D. R., Alvey, B., and Robbins, J. M. (1989). A direct manipulation interface for boolean information retrieval via natural language query. In *Proceedings of the 13th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '90, page 135–150, New York, NY, USA. Association for Computing Machinery.

Bartram, L., Correll, M., and Tory, M. (2021). Untidy data: The unreasonable effectiveness of tables. *IEEE Trans Vis Comput Graph*, 28(1):686–696.

Bendix, F., Kosara, R., and Hauser, H. (2005). Parallel Sets: visual analysis of categorical data. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 133–140. IEEE.

Bostock, M., Ogievetsky, V., and Heer, J. (2011). D^3: Data-driven documents. *IEEE Trans Vis Comput Graph*, 17(12):2301–2309.

Davies, J. (2012). Parallel Sets. `https://www.jasondavies.com/parallel-sets/`. Accessed Jaunary 12, 2024.

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *Journal of Statistics Education*, 3(3).

Dennig, F. L., Fischer, M. T., Blumenschein, M., Fuchs, J., Keim, D. A., and Dimara, E. (2021). Parsetgnostics: Quality metrics for Parallel Sets. In *Comput Graph Forum*, volume 40, pages 375–386. Wiley Online Library.

Dimara, E., Bezerianos, A., and Dragicevic, P. (2017). Conceptual and methodological issues in evaluating multidimensional visualizations for decision support. *IEEE Trans Vis Comput Graph*, 24(1):749–759.

Fischer, M. T., Frings, A., Keim, D. A., and Seebacher, D. (2021). Towards

a survey on static and dynamic hypergraph visualizations. In *2021 IEEE visualization conference (VIS)*, pages 81–85. IEEE.

Friendly, M. (1994). Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200.

Friendly, M. (1999). Extending mosaic displays: Marginal, conditional, and partial views of categorical data. *Comput Graph Stat*, 8(3):373–395.

Friendly, M. and Meyer, D. (2015). *Discrete data analysis with R: visualization and modeling techniques for categorical and count data*, volume 120. CRC Press.

Furmanova, K., Gratzl, S., Stitz, H., Zichner, T., Jaresova, M., Lex, A., and Streit, M. (2020). Taggle: Combining overview and details in tabular data visualizations. *Information Visualization*, 19(2):114–136.

Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H., and Streit, M. (2013). LineUp: Visual analysis of multi-attribute rankings. *IEEE Trans Vis Comput Graph*, 19(12):2277–2286.

Greenacre, M. (2017). *Correspondence analysis in practice*. CRC press.

Hartigan, J. A. and Kleiner, B. (1981). Mosaics for contingency tables. In *Computer science and statistics: Proceedings of the 13th symposium on the interface*, pages 268–273. Springer.

Healey, C. G. (1996). Choosing effective colours for data visualization. In *Proceedings of Seventh Annual IEEE Visualization'96*, pages 263–270. IEEE.

Hearst, M. A. (1999). User interfaces and visualization. *Modern information retrieval*, pages 257–323.

Hofmann, H. (2000). Exploring categorical data: Interactive mosaic plots. *Metrika*, 51:11–26.

Hofmann, H. (2006). *Multivariate Categorical Data — Mosaic Plots*, pages 105–124. Springer New York, New York, NY.

Hofmann, H. and Vendettuoli, M. (2013). Common angle plots as perception-true visualizations of categorical associations. *IEEE Trans Vis Comput Graph*, 19(12):2297–2305.

Im, J.-F., McGuffin, M. J., and Leung, R. (2013). GPLOM: The generalized plot matrix for visualizing multidimensional multivariate data. *IEEE Trans Vis Comput Graph*, 19(12):2606–2614.

Johansson Fernstad, S. and Johansson, J. (2011). A task based performance evaluation of visualization approaches for categorical data analysis. In *2011 15th International Conference on Information Visualisation*, pages 80–89. IEEE.

Kolatch, E. and Weinstein, B. (2001). CatTrees: Dynamic visualization of

categorical data using treemaps. https://cat-vis.github.io/src/data/papers_pdf/kolatch2001cattrees.pdf.

Kosara, R., Bendix, F., and Hauser, H. (2006). Parallel Sets: Interactive exploration and visual analysis of categorical data. *IEEE Trans Vis Comput Graph*, 12(4):558–568.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). UpSet: Visualization of intersecting sets. *IEEE Trans Vis Comput Graph*, 20(12):1983–1992.

Lomuscio, M. (2020). Sleep study. https://www.kaggle.com/datasets/mlomuscio/sleepstudypilot. Accessed January 11, 2024.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions On Graphics (Tog)*, 5(2):110–141.

Mauri, M., Elli, T., Caviglia, G., Uboldi, G., and Azzi, M. (2017). RAW-Graphs: A visualisation platform to create open outputs. In *Proceedings of the 12th biannual conference on Italian SIGCHI chapter*, pages 1–5.

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Muth, L. C. (2021). When to use quantitative and when to use qualitative color scales. Accessed January 11, 2024.

Nguyen, C. H. and Mamitsuka, H. (2020). Learning on hypergraphs with sparsity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2710–2722.

Peña-Araya, V., Xue, T., Pietriga, E., Amsaleg, L., and Bezerianos, A. (2022). HyperStorylines: Interactively untangling dynamic hypergraphs. *Information Visualization*, 21(1):38–62.

Pister, A., Prieur, C., and Fekete, J.-D. (2023). ComBiNet: Visual query and comparison of bipartite multivariate dynamic social networks. In *Comput Graph Forum*, volume 42, pages 290–304. Wiley Online Library.

Rao, R. and Card, S. K. (1994). The Table Lens: Merging graphical and symbolic representations in an interactive focus+context visualization for tabular information. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 318–322.

Reza, R. M. and Watson, B. A. (2019). Hi-D maps: An interactive visualization technique for multi-dimensional categorical data. In *2019 IEEE Visualization Conference (VIS)*, pages 216–220.

Rocha, M. M. N. and da Silva, C. G. (2018). Heatmap matrix: A multidimensional data visualization technique. In *Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI)*.

Russell-Rose, T. and Gooch, P. (2018). 2dSearch: A visual approach to search strategy formulation. In *Proceedings of the 1st Biennial Conference on Design of Experimental Search and Information Retrieval Systems*.

Schlimmer, J. (1987). Mushroom. UCI Machine Learning Repository. `https://doi.org/10.24432/C5959T`.

Schmidt, M. (2006). Der einsatz von sankey-diagrammen im stoffstrommanagement. Technical report, Beiträge der Hochschule Pforzheim.

Schonlau, M. (2003). Visualizing categorical data arising in the health sciences using hammock plots. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*.

Sedlmair, M., Meyer, M., and Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Trans Vis Comput Graph*, 18(12):2431–2440.

Shneiderman, B. (1994). Dynamic queries for visual information seeking. *IEEE software*, 11(6):70–77.

Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE symposium on visual languages*, pages 336–343. IEEE.

Spoerri, A. (1995). *InfoCrystal, a visual tool for information retrieval*. PhD thesis, Massachusetts Institute of Technology.

Symanzik, J., Friendly, M., and Onder, O. (2019). The unsinkable Titanic data. In *2019 Joint Statistical Meetings (ASA) Conference Proceedings*, Denver, USA. [Online]. Available: `https://www.datavis.ca/papers/JSM-2019-proceedings-final.pdf`.

Tenenhaus, M. and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50:91–119.

Theus, M. (2002). Interactive data visualization using Mondrian. *Journal of Statistical Software*, 7:1–9.

Theus, M. (2012). Mosaic plots. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):191–198.

Trye, D. (2022). Visualising multivariate categorical data. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, Tsukuba, Japan.

Trye, D., Apperley, M., and Bainbridge, D. (2023). Extending the Heatmap Matrix: Pairwise analysis of multivariate categorical data. In *2023 27th International Conference Information Visualisation (IV)*, pages 29–36.

Unwin, A., Hawkins, G., Hofmann, H., and Siegl, B. (1996). Interactive

graphics for data sets with missing values—MANET. *Comput Graph Stat*, 5(2):113–122.

Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N., and Fekete, J.-D. (2021). Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE Trans Vis Comput Graph*, 27(1):1–13.

Ware, C. (2019). *Information visualization: perception for design*. Morgan Kaufmann.

Wickham, H. and Hofmann, H. (2011). Product plots. *IEEE Trans Vis Comput Graph*, 17(12):2223–2230.

Wolfe, J. M. and Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):0058.

Young, F. W. and Bann, C. M. (1996). ViSta: The visual statistics system. Technical report, 94–1 (c), UNC LL Thurstone Psychometric Laboratory Research Memorandum.

Zeileis, A., Meyer, D., and Hornik, K. (2007). Residual-based shadings for visualizing (conditional) independence. *Comput Graph Stat*, 16(3):507–525.

Zhang, C., Chen, Y., Yang, J., and Yin, Z. (2019). An association rule based approach to reducing visual clutter in Parallel Sets. *Visual Informatics*, 3(1):48–57.

Zhang, Z., McDonnell, K. T., Zadok, E., and Mueller, K. (2014). Visual correlation analysis of numerical and categorical data on the correlation map. *IEEE Trans Vis Comput Graph*, 21(2):289–303.

Part III

Case Studies from te Reo Māori and New Zealand English

Chapter 6

Harnessing Indigenous Tweets: The Reo Māori Twitter Corpus

Having introduced the main visualisation techniques of interest in Part II, we now turn our attention to developing a corpus of Māori-language tweets, which will be useful for applying these techniques in a local linguistic context. We chose to focus on Twitter data because there was—at the time this research was carried out—a large community of Māori-language tweeters whose posts could be easily retrieved. This chapter details the procedure followed to collect and clean the data, and provides a preliminary analysis of frequent words, hashtags and users in the corpus. Our analysis is supported by numerous visualisations of different aspects of the RMT Corpus; however, unlike in previous chapters, these are explanatory rather than exploratory in nature, and represent a mixture of data types. The steps taken to clean the RMT Corpus are particularly relevant as they correspond to the early stages of the visualisation pipeline (how the data are transformed prior to being visualised) and have downstream effects for our case study of grammatical possession in Chapter 8.

Publication Details

The following paper[1] has been reproduced with minor changes to the formatting, as discussed in Section 1.4. Additionally, to ensure consistency with the rest of the thesis, all instances of "Sect." and "Fig." have been changed to "Section" and "Figure", respectively.

Trye, D., Keegan, T. T., Mato, P., & Apperley, M. (2022). Harnessing Indigenous Tweets: The Reo Māori Twitter corpus. *Language resources and evaluation*, *56*(4), 1229-1268. https://doi.org/10.1007/s10579-022-09580-w

Abstract

Te reo Māori, the Indigenous language of Aotearoa New Zealand, is a distinctive feature of the nation's cultural heritage. This paper documents our efforts to build a corpus of 79,000 Māori-language tweets using computational methods. The *Reo Māori Twitter (RMT) Corpus* was created by targeting Māori-language users identified by the *Indigenous Tweets* website, pre-processing their data and filtering out non-Māori tweets, together with other sources of noise. Our motivation for creating such a resource is three-fold: (1) it serves as a rich and unique dataset for linguistic analysis of te reo Māori on social media; (2) it can be used as training data to develop and augment Natural Language Processing (NLP) tools with robust, real-world Māori-language applications; and (3) it will potentially promote awareness of, and encourage positive interaction with, the growing community of Māori tweeters, thereby increasing the use and visibility of te reo Māori in an online environment. While the corpus captures data from 2007 to 2020, our analysis shows that the number of tweets in the RMT Corpus peaked in 2014, and the number of active tweeters peaked in 2017, although at least 600 users were still active in 2020. To the best of our knowledge, the RMT Corpus is the largest publicly-available collection of social media data containing (almost) exclusively Māori text, making it a useful resource for language experts, NLP developers and Indigenous researchers alike.

[1] Haehae ana te whatumanawa mōu e te rangatira, e te hoa. Moe mai. We acknowledge the sudden passing of one of the authors, Paora Mato. Rest in peace.

6.1 Introduction

"Ka ngaro te reo, ka ngaro tāua, pērā i te ngaro o te Moa."
If the language is lost, man will be lost, as dead as the Moa.
(Waitangi Tribunal, 1986, p. 7).

Te reo Māori[2] (the Māori language) is intrinsic to Māori culture, and constitutes an important part of New Zealand's national identity. In spite of this, the Māori language is currently endangered, largely due to the ongoing effects of colonial contact and the resultant marginalisation of the Māori people, their language and their culture. While significant effort has been made to revitalise the Māori language over the past 30–50 years (King, 2018; Harlow and Barbour, 2013; Harlow, 2007), substantial work remains "to mitigate ongoing language shift and loss for Māori ... in Aotearoa/New Zealand" (May and Hill, 2018, p. 317). This paper documents efforts to compile a corpus of Māori-language tweets, which we call the *Reo Māori Twitter Corpus* (hereafter, the *RMT Corpus*). The RMT Corpus consists of 79,018 original tweets, comprising just over one million words and capturing output from 2302 accounts. It is anticipated that the creation of this resource, and any others derived from it in the future, will enhance the study, status, use and, therefore, the health of te reo Māori in wider contexts.

6.1.1 Roadmap

The structure of the paper is as follows. Section 6.2 provides important background information and explains our motivation for this work. In Section 6.3, we describe related work, focusing on existing Māori corpora and Natural Language Processing (NLP) resources, as well as highlighting issues of data sovereignty from an Indigenous Māori perspective. Section 6.4 details the method for building the RMT Corpus, summarising the process in four main steps. A preliminary analysis of the RMT Corpus, which we plan to extend in future work, is given in Section 6.5. Section 6.6 explains how researchers can download the RMT Corpus and associated resources. Finally, in Section 6.7, we present conclusions and suggestions for future work.

[2]In this paper, the terms *Māori*, *te reo* and *te reo Māori* are used interchangeably. We use macrons to denote long vowel sounds, following Te Taura Whiri's (2012) *Guidelines for Māori Language Orthography.*

6.2 Motivation

Modern NLP resources are usually developed using quantitative methods that rely on the existence of large corpora. Unfortunately, substantial corpora for minority languages (such as Māori) are scarce; they exist mostly for languages that have a market-driven need for NLP tools, such as English, French, German and Cantonese. In fact, it has been estimated that significant digital resources exist for only 20–30 of the world's 6000+ languages (Maxwell & Hughes, 2006, as cited in Bloem et al., 2019; see also Scannell, 2007). As an endangered language spoken by roughly only four per cent of the New Zealand population,[3] it is not surprising that Māori falls outside this category. A lack of appropriate corpora hampers the development of critical NLP resources for the Māori language (James et al., 2020). We suggest that a corpus of Māori tweets—while no panacea—is a step in the right direction for future progress in NLP for te reo Māori.

The internet and the proliferation of social media provide unprecedented access to many low-resource languages. Large amounts of untapped, real-time data are available on popular social media sites, such as Twitter, Facebook and Instagram, creating an ideal hunting ground for computational linguists (King, 2015). Anecdotal evidence suggests that the Māori language is increasingly being used and discussed on these platforms, as greater numbers of New Zealanders endeavour to incorporate te reo Māori in their everyday lives. Such outcomes are the fruit of a wide range of language revitalisation initiatives aimed at promoting the use of te reo Māori in everyday contexts. For instance, a national social media campaign was launched by Te Māngai Pāho (the Māori Broadcast Funding Agency) in September 2019 to achieve a target of one million te reo Māori tweets by New Zealanders, using the hashtags *#MahuruMaori*, *#1MirionaTihau* and *#1MirionaTweets* (Mereraiha, 2019).

Corpora derived from social media have a diverse range of applications. They can be used to build language models, train word embeddings and perform various kinds of machine learning and NLP tasks, such as part-of-speech tagging, morphological analysis, parsing, information retrieval, speech recognition and machine translation (see, for example, El-haj et al., 2015). This is particularly important for languages such as Māori, because the creation of new NLP resources for endangered languages also enhances their profile, contributing to more positive attitudes towards the language, and a sense of greater relevance and modernity. This in turn attracts new learners (Coto Solano

[3]The 2018 New Zealand census reports a total of 185,955 Māori speakers (Stats NZ, 2020).

et al., 2018) and increases language prestige (Bender, 2019; Meyerhoff, 2019).

Social media corpora are also gaining prominence in linguistics. In the context of minority languages, these corpora are especially useful when limited data are available elsewhere, or when existing resources suffer from a lack of currency and diversity. Twitter provides a source of natural, user-generated content that is very different from other, more traditional genres. Cassels (2019) notes that social media sites are conducive to spontaneous literacy production and non-prescriptivist language use, reflective of everyday usage. As a result, linguistic phenomena rarely seen in formal settings are more likely to be present on Twitter(Lynn and Scannell, 2019). Moreover, since tweets often resemble spoken language more closely than formal written language, they are more likely to be suitably used as training data for text-to-speech and speech-to-text applications.

Research also indicates that social media and the internet can play an important role in supporting the revitalisation of minority languages. This is in keeping with traditional language revitalisation theory (see, for example, Giles et al., 1977 and Zuckermann, 2020), which maintains that the more domains in which a language is used, the stronger its *ethnolinguistic vitality*. Languages with strong ethnolinguistic vitality are visible and accepted within their communities and, consequently, more likely to be transmitted to future generations (Meyerhoff, 2019). Cunliffe et al. (2013, p. 75) found that social media sites help to establish "linguistic communities" and "revive weakened languages". In a similar vein, Ka'ai (2017, p. 30) observes that "minority-language users can, through the internet and new media, become producers as well as consumers of media products in their own language" (see also Ka'ai et al., 2012). This is not to say that the use of a minority language on social media and the internet will *prevent* its extinction; there are many other important contributing factors, such as the number of speakers, the level of institutional support that the language receives and the strength of intergenerational transmission (Bird, 2020). However, young people's importance in language survival and their increasing comfort and dependence on social media communications cannot be overlooked (Keegan and Cunliffe, 2014).

From a Māori perspective, there are a range of benefits and drawbacks concerning the uptake of social media. Sciascia (2016, p. 6) notes that social media sites can function as a "virtual *marae* [gathering space]" where the Māori language is "becoming a normalised form of communication" and "*tikanga* [customs/values] are being practiced". In particular, social media aligns with the Māori values of *tino rangatiratanga* (self-determination) and

whanaungatanga (relationships/networks). The term *e-whanaungatanga* (electronic relationships/networks) was coined in recognition of the implications of social media for tikanga Māori, to capture the notion of cultivating positive relationships online (Waitoa et al., 2015).

As regards disadvantages, social media increases susceptibility to cultural appropriation and can be used as a platform for targeting ethnic minorities and/or spreading racist views and disinformation——all of which clearly oppose Māori values. Furthermore, the widespread use of artificial intelligence (AI) on digital platforms poses ethical concerns relating to algorithmic bias, discrimination, privacy (Bender et al., 2020) and especially data sovereignty (Whaanga, 2020; see Section 6.3.2). Another drawback of the use of social media is the impact of diaspora and the fact that Māori are less inclined to travel back to their *tūrangawaewae* (domicile of origin) because they are connected through social media. However, at the same time, social media does enable the displaced or diasporic to continue using their language, and to maintain their connections to their homelands and each other.

The benefits and drawbacks of the use of te reo Māori on social media platforms are part of a much larger discussion; however, we believe there is significant value in creating a corpus of Māori-language tweets. Twitter has been selected in favour of other social media platforms, due to the unidirectional nature of its network and the accessibility of data through the Twitter API. For the most part, Twitter users are able to follow other users without those users' permission. The same is true for developers wanting to download their tweets. We are confident, therefore, that Māori-language data can be more readily extracted from Twitter than from other sites such as Facebook, whose privacy settings severely limit the amount of data that can be collected. Targeting the Twitter platform enables the creation of a larger corpus than we might get from other platforms. Additionally, the aforementioned Te Māngai Pāho campaign was specific to Twitter, and that significant amount of generated data is not available elsewhere.

The RMT Corpus lends itself to linguistic analysis of informal Māori language as it captures authentic, user-generated content from real (human) tweeters of te reo. Furthermore, the corpus offers great potential for contributing to the development of new and important NLP resources for the Māori-language community. For instance, it could be used as a golden set (benchmark) of Māori-annotated tweets for language classification purposes, as training data for text-to-speech and speech-to-text applications, or as dictionary input for a reo Māori auto-completion tool (Innes, 2021). To this end,

the corpus has already been shared with an iwi media organisation to support their language revitalisation efforts, as well as several research institutes across New Zealand. Through the networks of the Māori authors of this paper, the corpus will also be shared with Māori-language teachers who are interested in using real language examples and creating fresh learning materials for their students.

6.3 Related Work

The following is a summary of the resources and techniques that informed our approach to building a corpus of Māori-language tweets. The present study builds on the findings of two prior investigations into Māori language use on Twitter. Mato and Keegan (2013) analysed the ten most prolific users of te reo Māori and found that they tweeted primarily to disseminate religious text, news items or organisational notices, rather than to present their own thoughts and opinions, or engage interactively with other users (p. 190). In a follow-up study, Keegan et al. (2015) found that a small yet dedicated group of individuals on Twitter was actively conversing in te reo Māori, especially during events with strong cultural significance, such as Māori Language Week[4] and Te Matatini, a nation-wide Māori performing arts festival and competition. This paper serves as an update to these studies, whilst also collating a Māori-language corpus: a first-of-a-kind resource made available to other researchers. A defining feature of the RMT Corpus is the enriched metadata associated with both the tweets and their authors, which affords new possibilities for exploring Māori-language use on social media, especially in terms of content and location (however, such exploration is beyond the scope of the present study).

6.3.1 Existing Māori-language resources

Several Māori-language corpora have been compiled over the past few decades. These resources typically capture more formal language than one would expect to see on Twitter, arising from genres such as legal texts and speeches, television broadcasts and children's literature. Table 6.1 provides a summary of each corpus, citing information about their size, genre and availability.

[4]This is an annual government-sponsored initiative, now officially called *Te Wiki o te Reo Māori*.

Table 6.1: Summary of existing Māori-language corpora.

Title	Date	Description	Medium	Publicly Available	Source
Legal Māori Corpus (LMC)	1829-2009	8 million words from legal and law-related texts; the largest structured corpus of the Māori language.	Written	Partially: Only data from before 1910 (5.3m words) can be downloaded, but the entire corpus is searchable online	Boyce (2011; 2016; 2022)
Niupepa Collection	1842-1932	Over 17,000 pages of historic newspaper text, based on a microfiche collection produced by the Alexander Turnbull Library. 70% of the collection is written exclusively in Māori.	Written	Yes	Apperley et al. (2002), NZDL (2002)
Hansard Corpus	1860-2018	573,000 words from NZ Parliament utterances that are predominantly in Māori.	Spoken	Yes	Te Hiku Media (2021a)
MAONZE Corpus	Late 1940s; 2001-2009	Transcribed recordings of the Māori and English speech of three generations of Māori (109 hours, 620,700 words).	Spoken	No	King et al. (2010)
Tūhoe Corpus	2009	Transcribed recordings of five male and five female elders of the Tūhoe tribe (19 hours, 114,000 words).	Spoken	No	King et al. (2010)
Māori Broadcast Corpus (MBC)	1995-1996	Transcripts of one million words of radio and television broadcasts.	Spoken	No	Boyce (2006)
Māori Texts for Children (MTC)	Unknown	A collection of Māori texts for children.	Written	No	Huia Publishers
Te Paipera Tapu	2012 (first translated 1868)	A one-million word Māori translation of the Bible, reformatted to include macrons.	Written	Yes: Users can download the Te Paipera Tapu app, or view the text online	Bible Society New Zealand (2021)
Māori Written Corpus	2009-2012	3.7 million words of Māori text, sourced from personal collections, language friends, public and government sites (personal communication, February 24, 2021).	Written	Yes: On request	Te Taka Keegan
Māori Speech Corpus	2019-2020	A Māori speech corpus, sourced from Ngā Mahi a Ngā tūpuna (Grey, 1928), together with a lexicon of 10,000 words and 18,000 names, and 1,030 recorded sentences.	Spoken	No	James et al. (2020)
Papa Reo Dataset	2017-2020	A dataset compiled by Te Hiku Media, based on 198,000 speech recordings from 2,200 speakers, comprising 4.8 million words and 5,000 unique sentences.	Spoken	No, but see the resulting tools (Te Hiku Media, 2021c)	Moses et al. (2020)

The RMT Corpus complements this existing body of Māori-language text since, although a mixed-language corpus of three million English tweets exists that contains borrowed Māori words or loanwords (Trye et al., 2019, 2020), there is, to the best of our knowledge, no social media corpus comprising (almost) exclusively Māori-language text.

Te reo Māori has more recently received considerable attention from the NLP community (more so than other Polynesian languages; see Coto Solano et al., 2018). For instance, te reo Māori is supported by the popular machine translation tool *Google Translate* (Google, 2006), an online tool exists for automatically restoring Māori macrons (Cocks and Keegan, 2014), and SwiftKey has developed a custom Māori keyboard. Other initiatives have yielded the first automatic Māori speech recogniser and transcription tool, recently developed by Te Hiku Media (2021c), and the use of deep learning methods to provide instant, character-by-character pronunciation feedback to learners of te reo Māori (Moses et al., 2020). In a related project, a synthetic male voice was developed, allowing synthesised speech to be generated from any Māori text input (Shields et al., 2019). Furthermore, Te Hiku Media is in the process of developing the first reo Māori part-of-speech (POS) tagger (Finn, 2021). Despite these and other similar developments, advancements for te reo Māori, in terms of the data and equipment that are required to build robust speech technologies and language models, remain woefully under-resourced (James et al., 2020).

6.3.2 Data sovereignty

Indigenous communities are becoming increasingly aware of the importance of maintaining control and sovereignty over their data, and of the potential repercussions of not being able to do so (Kukutai and Taylor, 2016). Te Hiku Media advocates a moral responsibility to "respect the *mana* [honour/power/respect] of the data and the people from whom it descends" (Te Hiku Media, 2021b). Unfortunately, this is not always straightforward. In today's culture of open data and open science, the sharing of information can perpetuate the misuse and subsequent exploitation of data for commercial gain—more especially by large corporations (even if this is not the intention of the people who collect and share that data). Even so, Indigenous people can benefit hugely from cutting-edge contemporary technologies, particularly in terms of currency and scalability. However, they can also be exposed to further marginalisation and data appropriation; a form of modern colonisation in the digital space (Keegan et al., 2021). Bird (2020) calls for speech and language technologists to con-

sider their own role in commodifying Indigenous knowledge as data, cautioning against techno-solutionism in NLP. In terms of this research, we believe it is appropriate to make the RMT Corpus widely available as a resource that has been sourced from an open platform, which is intended to be used by academics and Indigenous researchers to benefit Māori and the Māori-language community.[5]

6.3.3 Welsh Twitter corpus

While there have been several exciting developments in the world of Māori NLP (see Section 6.3.1), none of these resources makes use of the rich and ever-increasing amount of data available on social media. The *Welsh Twitter Corpus* (Jones et al., 2015) demonstrates the advantages of building a Twitter corpus for a minority language. Developed at Bangor University, the corpus comprises seven million Welsh-language tweets, and helps to meet "great demand" for informal Welsh data (ibid). The authors identify several potential applications for the *Welsh Twitter Corpus*, which we believe are also relevant to the RMT Corpus in the context of te reo:

- Training predictive text systems for [sms on] phones
- Finding new words in the Welsh [language]
- Developing research material for academic departments in Universities, including in linguistic, sociological and medical studies
- Educational and demonstration data for children and young people in coding clubs
- Valuable information for the market, for example; tracking and analysing user emotions (sentiment analysis).[6]

6.4 Building the RMT Corpus

In this section, we explain our methodology for building the RMT Corpus. Our main objective was to create a corpus of (almost) exclusively Māori-language tweets. Initially, we tried to mirror the approach employed by Trye et al. (2019), which used a seed list of Māori *query words* to build a corpus of New Zealand English tweets (English tweets containing Māori borrowings). To this end, we compiled a set of Māori words and phrases that we considered would be useful for harvesting tweets explicitly tagged as Māori. However, it soon

[5]We expect researchers to adhere to the international data principles of FAIR (Findable, Accessible, Interoperable and Reusable) and CARE (Collective Benefit, Authority, Responsibility and Ethics).

[6]Bullet points copied directly from Jones et al. (2015).

became clear that this approach would not work as intended, because Twitter does not provide official support for te reo Māori. This meant that we could not automatically extract Māori-language tweets using only the Twitter API.

Instead, we decided to gather (and clean) data from known *users* of te reo. Targeting a set of relevant users is a well-established method for compiling a Twitter corpus (see, for instance, Verhoeven et al., 2016 and Zaghouani & Charfi, 2018). Māori-language tweeters were identified using the *Indigenous Tweets* website (Scannell, 2011a), which periodically crawls the web for tweets written in 185 minority languages. Since our approach relies heavily on *Indigenous Tweets*, we provide a brief overview of its functionality in Section 6.4.1, and describe relevant limitations in Section 6.4.2.

We now provide an overview of how the RMT Corpus was created, before explaining each step in detail. The corpus-building process can be summarised in four main steps, as shown in Figures 6.1 and 6.2. In step one, the *Twitter API* (Twitter, 2021c) was used to download 11 million tweets from Māori users listed on the *Indigenous Tweets* website (Section 6.4.2). Because users tend to post much more frequently in English than Māori, only a small portion of these tweets were written in te reo.

Step two involved pre-processing the data to enhance readability and mitigate noise in the corpus. At this point, 1.5 million short tweets were discarded, along with 60,000 tweets from automated accounts. Tweet text was also formatted such that @user mentions and hyperlinks were standardised. Two distinct versions of the text were extracted: one in which special characters such as emojis were preserved, and the other in which these were eliminated (Section 6.4.3).

In step three, the percentage of Māori text in each tweet was calculated, so that tweets containing less than 70–80% Māori text could be systematically removed. 9.4 million non-Māori tweets were filtered out accordingly. This step—crucial for cleaning the corpus—was carried out by leveraging NLP code developed by Te Hiku Media (Section 6.4.4).

Finally, step four involved removing 4473 formulaic Māori tweets that were thought to adversely affect corpus quality (Section 6.4.5). The remaining tweets in the corpus are believed to contain original and authentic Māori text, suitable for a range of computational and linguistic tasks, as described in Section 6.2. It is worth noting that even the discarded tweets can be seen as a valuable resource, because they represent varying degrees of Māori/English code-switching (captured by the *percent_maori* attribute; Section 6.4.4).

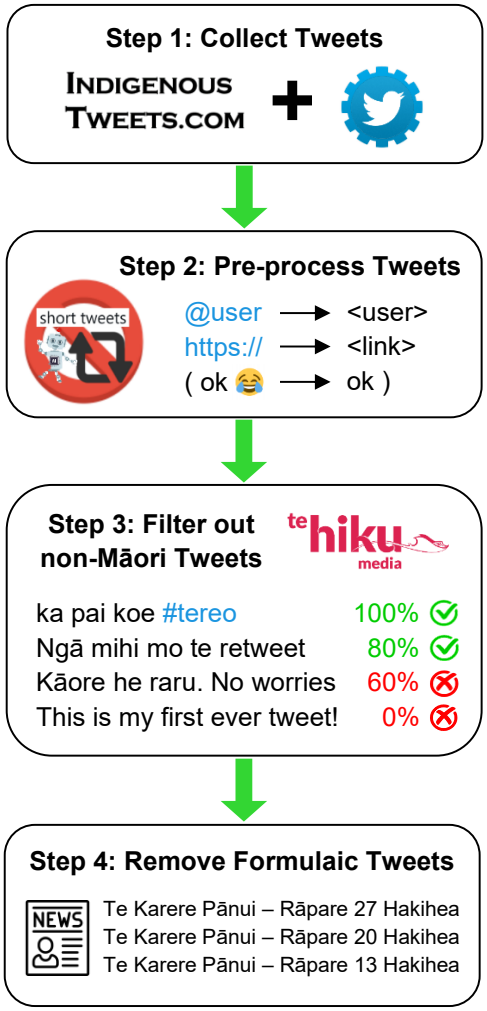


Figure 6.1: The process of building the RMT Corpus, broken down into four key steps.

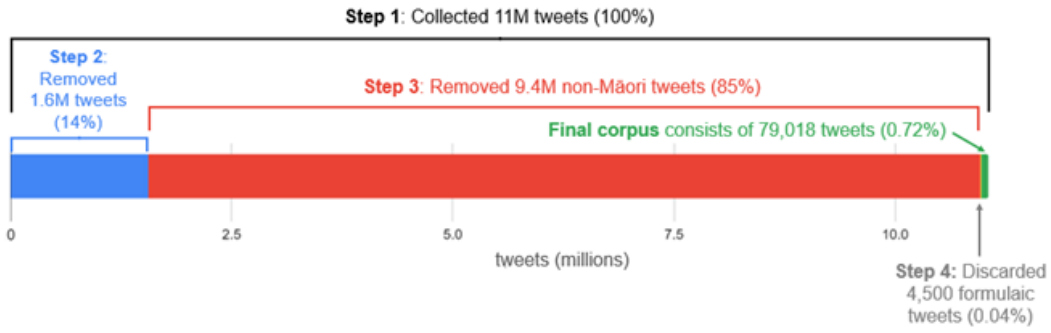


Figure 6.2: Visualisation showing the proportion of tweets removed during each step. The width of the entire bar represents the 11 million tweets that were initially collected, and the green bar on the right represents the final corpus of 79,018 tweets. The vast majority of tweets (85%) were removed during Step 3, after calculating the percentage of Māori text in each tweet.

6.4.1 Indigenous Tweets background

The *Indigenous Tweets* website catalogues users of 185 Indigenous languages (as of 8 February, 2021), including te reo Māori. Created by Professor Kevin Scannell of Saint Louis University in March 2011, the website aims to connect minority-language speakers through the internet, making their voices visible in the face of language giants such as English and Spanish (Scannell, 2011b). *Indigenous Tweets* uses a combination of character tri-grams and word features to determine whether a given tweet belongs to a particular language in its database. A separate crawler exists for each language, which scans the corresponding list of users and keeps all tweets detected in that language. The database is continually updated to not only incorporate new tweets from known users, but to also capture other (previously unidentified) users of the target language, including their older tweets. More specifically, there are four main techniques for adding tweets from new (Māori) users (for further details, see Scannell, 2022):

1. Leveraging distinctive (Māori) words not present in other languages;
2. Searching accounts that known (Māori) users have retweeted;
3. Querying known (Māori) users' followers;
4. Crowdsourcing suggestions for (Māori) users via the *Indigenous Tweets* website.

The *Indigenous Tweets* home page reports key statistics for all 185 featured languages, including the total number of tweets, distinct users, the most frequent user and the user who posted the first recorded tweet for each language. When sorted by number of tweets, Māori is currently ranked sixteenth overall, with 313,122 tweets in total.[7] Interestingly, however, Māori rises to fifth place when sorted by number of users, with what appears to be a healthy community of 3,090 users. This would suggest that, compared to other Indigenous languages with a large presence on Twitter, Māori has *more users* who post *fewer tweets*. These numbers have increased significantly over the past six years; in 2015, there were only 93,283 Māori tweets posted by 345 tweeters (Keegan et al., 2015).

Users of *Indigenous Tweets* can also view more detailed information about each language's top 500 tweeters. Figure 6.3 shows the top 20 Māori tweeters, three quarters of whom were active within the past year. Most Māori-language users only tweet in te reo a small proportion of the time, presumably tweeting

[7]The RMT Corpus comprises only a quarter of these tweets, for reasons explained in Sections 6.4.2 and 6.4.4.

Māori	katoa	% Māori	apataki	whakaapataki	He Tweet mātāmuri
29810	32020	93.1	21	0	2013-05-16 13:44:42
22390	24659	90.8	17	0	2012-06-21 19:04:02
11646	144401	8.1	6554	4043	2021-01-31 19:03:02
11620	12507	92.9	12	0	2013-06-05 08:33:03
10689	19865	53.8	3732	2801	2021-01-31 21:42:48
5065	128094	4.0	1640	2620	2021-01-31 15:15:22
3611	6952	52.0	5348	386	2021-01-28 03:24:21
2927	87548	3.3	9550	1282	2021-01-31 18:32:21
2892	13111	22.1	812	980	2019-12-04 03:30:39
2728	31330	8.7	1735	2277	2021-01-31 20:35:58
2697	65481	4.1	3343	1216	2021-01-31 18:34:26
2628	26420	9.9	6729	543	2021-02-01 00:48:39
2542	15034	16.9	297	0	2021-02-01 00:39:36
2509	5007	50.1	1751	299	2021-01-31 20:35:09
2474	49979	5.0	2741	4665	2021-02-01 00:37:41
2369	3707	63.9	1024	876	2020-10-19 18:57:07
2367	19968	11.9	1097	1471	2021-01-16 12:41:35
2156	49025	4.4	731	150	2015-01-31 17:31:52
2058	12228	16.8	2965	833	2021-01-31 23:44:33
1794	2451	73.2	482	613	2017-05-31 05:01:21
1774	78748	2.3	5405	2431	2021-01-31 18:02:46

Figure 6.3: The top 20 Māori tweeters on the *Indigenous Tweets* website as of 8 February, 2021. Usernames have been omitted to protect users' identities.

mainly in English the rest of the time.[8] These figures also include retweets, a re-posting of a tweet, which can lead to overestimates of how many tweets have been written in a particular language, although this still provides evidence of people engaging with the content in that language.

The *Indigenous Tweets* website has three notable limitations that users intending to build a minority-language corpus should be mindful of. First, the web crawlers used are not perfect and occasionally admit false positives. We found a small number of non-Māori language users whose (Indonesian and Samoan) tweets were erroneously included in the database. Second, the process of updating the list of tweeters for a given language is not entirely automated, which means the website may be missing information for recently discovered users. Finally, the classifier does not currently support detection of multiple languages within the same tweet (e.g. due to code-switching). Assigning

[8]This is a reasonable assumption to make, given that all adult Māori speakers are bilingual, and English is the dominant language of New Zealand.

exactly one label to each tweet is not always representative of actual language use and may preclude bilingual users who frequently post content in two or more languages from being identified.

6.4.2 Step one: collecting tweets

As explained previously, the first step involved in creating the RMT Corpus was to collect tweets from known Māori-language users. Since the *Indigenous Tweets* website displays only the top 500 users for each language, we requested access to the usernames of all 3090 (potential) Māori tweeters (Kevin Scannell, personal communication, 21 January, 2020). The Twitter API was then used to gather as many tweets as possible from these users (including their non-Māori-language tweets). A total of 10,870,247 tweets were collected. Our search timeframe spanned a period of nearly 15 years, from Twitter's inception in March 2006 through to December 2020, although the first Māori tweet did not occur until May 2007. Not all tweets were available for every user, due to the corresponding accounts becoming protected (accessible only to the user and their followers), or having been suspended[9] or deleted (see Figure 6.4). Moreover, some users may have changed their usernames from those specified on the Indigenous Tweets website, resulting in a false association. In particular, we were not able to gather data for 286 of the 3,090 listed users in 2020 (9.26%), which can also be attributed to a change in how the tweets were collected.[10] We have augmented the tweets with a rich array of metadata (some sourced directly from the Twitter API, others manually derived), which we believe may provide useful insights into patterns of Māori-language use on Twitter.

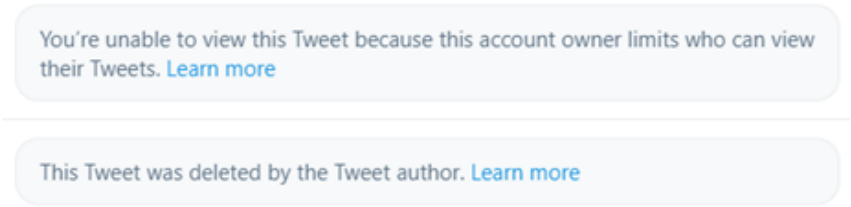


Figure 6.4: Searching for a tweet that is not publicly available on Twitter. The API returns a "Not Found Error" if a tweet has been deleted, or an "Authorization Error" if the tweet is from a protected account.

[9]Twitter suspends accounts that violate its rules (Twitter, 2021d)

[10]We previously had premium access to the Twitter API, which provided access to historical tweets from protected and deleted accounts up to the beginning of 2020, but we were not able to collect such tweets throughout 2020.

6.4.2.1 Tweet metadata

The RMT Corpus has a wealth of metadata available, with (up to) 48 variables pertaining to both the tweets and their authors. This includes a mixture of *upstream metadata*, sourced directly from the Twitter API, and *derived metadata*, which we added ourselves. Unfortunately, many records are incomplete; see Section 6.4.2.3.

A selection of tweet metadata included in the corpus is detailed below (see Appendix D for a description of all 48 variables):

- The text contained in the tweet, formatted as per Section 6.4.3. Users can choose whichever version of the text is best suited to their specific needs: the *content* without special characters, or the *content_with_emojis*, which includes special characters [derived].
- The tweet *ID*, which uniquely identifies the tweet and allows it to be viewed in context [upstream].
- The timestamp (*date*) when the tweet was posted, which can be used for diachronic analysis [upstream].
- The *conversation ID*, which is the ID of whichever tweet initiated the conversation (Twitter, 2021a) [upstream].
- The number of *likes*, *retweets* and *quotes* associated with the tweet; so-called "public metrics" [upstream]. These three values were added together to compute the number of *favourites* [derived].
- The number of *replies* that the tweet received [upstream].
- The *language* by which the tweet was classified. Because the Twitter API does not support te reo Māori, most tweets in the corpus have been erroneously classified as another language [upstream].
- The *source* from which the tweet was posted, including a wide range of devices and third-party applications [upstream].
- *Error* information explaining why some tweet data could not be retrieved (e.g. not found or unauthorised) [upstream].
- The list of *Māori words* detected in the tweet (during Step 3; see Section 6.4.4) [derived].
- The *percentage* of Māori text detected in the tweet (during Step 3; see Section 6.4.4) [derived].

6.4.2.2 User metadata

Each tweet is authored by a *user* who has their own set of attributes. The RMT Corpus comprises (up to) 26 attributes for 2,302 distinct users. This information is shared among all tweets by the same author, and can be rep-

resented in a nested data structure (embedded within the tweet metadata). The author metadata is time-dependent, providing an accurate snapshot from December 2020.

A selection of user attributes included in the corpus is given below (see Appendix D for the complete list):

- An *alias* for the user who posted the tweet. This is in the form *TX*, where *X* is the user's ranking based on their total number of tweets in the corpus, such that user *T1* is the most prolific Māori-language tweeter. We have anonymised usernames in this way to safeguard against malicious use of the data [derived].
- The account's *status* as of April 2021 (active, protected, not found or suspended), which influences whether the tweet can (still) be downloaded via the Twitter API [upstream].
- The user's total number of *tweets* in the corpus [derived].
- The user's *location*, based on self-reported information. Where possible, this was aggregated into New Zealand regions (e.g. "Auckland", "Waikato") and names of overseas countries (e.g. "France", "Australia"). Many users did not specify their location, or did so at a higher granularity (e.g. "North Island"), which meant we could not infer the particular region where they live [derived].
- The date the user's account was *created*, which may be fruitfully compared with the date of the user's first Māori tweet in the corpus [upstream].
- The user's total number of *statuses*, including non-Māori tweets and retweets [upstream].
- The user's total number of *favourites*, calculated as the sum of all their likes, retweets and quoted tweets [upstream].
- The user's total number of *followers* [upstream].
- The number of accounts that the user follows (*friends*) [upstream].
- Whether or not the user's account is *verified* (this applies to accounts of public interest that are "authentic, notable, and active"; Twitter, 2021b) [upstream].

With a view to understanding more about user demographics, one of the authors of the paper coded additional information about each tweeter's gender and ethnicity. The tagging process involved manually reading each user's account description to see if they mentioned their ethnicity (e.g. "Māori") and/or personal pronouns (e.g. "she/her/ia"). In some cases, these were not explicitly stated but revealed implicitly. For instance, the coder inferred that

users were Māori if they mentioned their iwi/ tribes, and similarly deduced their gender if they used nouns such as *māmā* or "mother" (e.g. "proud māmā of two"). Accounts representing multiple people (e.g. businesses or organisations) were labelled as "groups". We were able to extract gender information for 85% of users (n = 1957), among whom 50% were female, 31% were male, 19% were group accounts and 0.3% were gender-neutral. This information reveals a richer picture of the social profile of Māori-language tweeters, which we intend to explore in future work.

6.4.2.3 Overview of missing metadata

While we attempted to collect data for 48 variables, most tweets are missing information for one or more of these fields. Figure 6.5 is a waterfall chart showing how many tweets have data recorded for the corresponding number of variables. All tweets comprise a minimum of 15 variables, including the tweet content, tweet ID, user alias, list of Māori words and percentage of Māori text. Only 11,354 tweets in the corpus (14.37%) are missing data for more than five of the 48 variables; the remaining 67,664 tweets constitute complete or almost complete records.

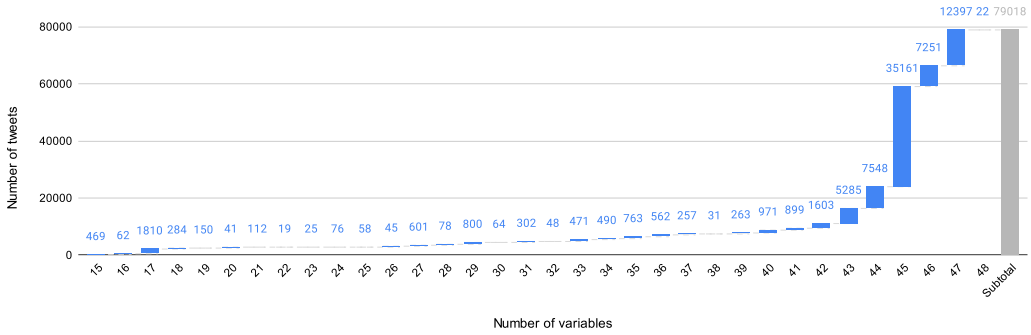


Figure 6.5: Waterfall chart showing how many tweets have data for the given number of variables. 85.63% of tweets have at least 43 of the 48 variables.

Figure 6.6 provides a more detailed visual summary of the missing information. Specifically, the rows in the matrix show all combinations of variables that occur more than 500 times, accounting for 84% of the corpus data. The most frequent combinations are listed at the top, reinforced by the shade of the (non-black) cells and the bar chart on the right-hand side. The columns provide an indication of how many tweets have metadata for the corresponding variable, with *in_reply_to_user_id*, *user.iwi* and *user.ethnicity* being the most sparse variables. Note that, in some cases, there is a distinction between blank (null) values and missing (unknown) values. For example, regarding the

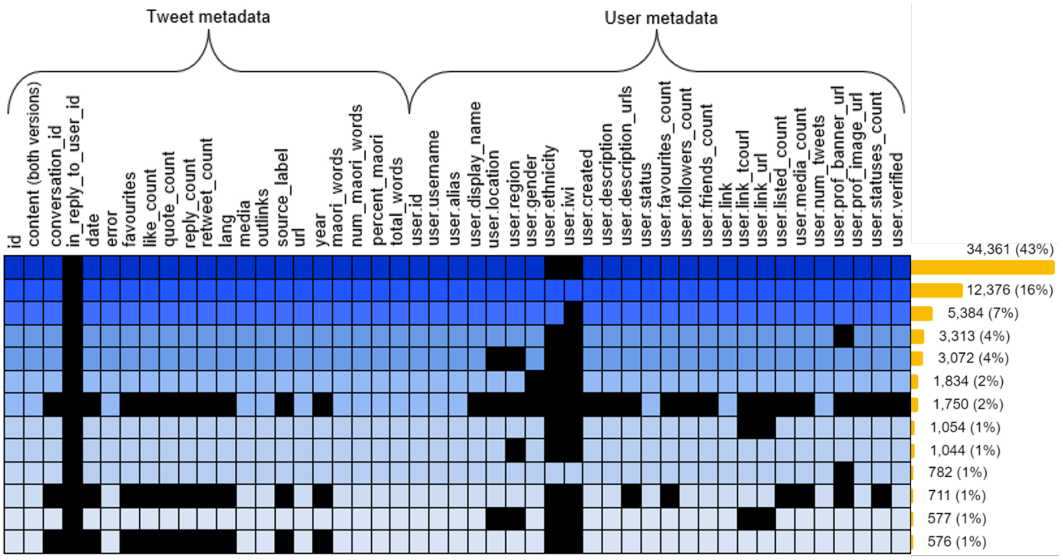


Figure 6.6: Matrix showing the most frequent *combinations* (rows) of tweet and user *metadata* (columns) in the RMT Corpus. Missing values are denoted using black cells. Drawn using Microsoft Word.

user.description variable, this would be the difference between a user who does not have an account description, and a user whose account description is not known, because we were not able to retrieve it. Values in the former category are not considered to be missing.

6.4.2.4 Corpus collection caveats

The RMT Corpus may not be fully representative of Māori language use on Twitter. As noted at the beginning of Section 6.4.2, various constraints imposed by Twitter and our data collection method prevented us from accessing a (potentially) large number of tweets posted by certain users, especially throughout 2020. In addition, the *Indigenous Tweets* database does not provide an exhaustive list of Māori tweeters (see Scannell, 2022), which means we cannot claim to have accounted for every single te reo Māori tweeter. Consequently, the RMT Corpus should be regarded as capturing only a subset of relevant tweets from a subset of relevant users.

In the interests of transparency and replicability, we describe some further caveats below. Twitter's terms of service do not allow third parties to redistribute tweet content directly: only IDs and selected metadata can be publicly shared. Anyone interested in accessing the corpus text is thus required to download ('hydrate') the tweets using the IDs provided, but this relies on their still being available via the Twitter API (Section 6.6). As of April 2021, 6,443 tweets in the RMT Corpus (8.15%) could not be hydrated, either because

the user's account was protected, suspended or deleted, or because individual tweets had been deleted from active accounts. Other tweets that are currently available may also become inaccessible in the future, exacerbating this problem (see further discussion in McCreadie et al., 2012).

Two key features of Twitter are the ability to reply to tweets and to package multiple related tweets in a conversation called a "thread". Therefore, because tweets are often part of larger conversations, important context may be missing when a tweet is read in isolation. While a *conversation_id* is available for most tweets in the RMT Corpus, enabling an analyst to group tweets by conversation, there is no guarantee that every tweet in a conversation will feature in the corpus. Not all tweets belonging to a given conversation will necessarily also be written in Māori, or contain sufficient Māori text. In addition, some tweets in a conversation may not be publicly available, harking back to the previous problem.

As explained in Section 6.4.2.3, some tweets are missing a considerable amount of metadata (see Figures 6.5 and 6.6). We assume that any self-reported user information (such as location and gender) is accurate. From a sociolinguistic perspective, the RMT Corpus was not designed to yield a balanced sample of users (in terms of variables such as region, gender and ethnicity), which affects how it can be used. Finally, analysts should be aware that each account in the RMT Corpus does not necessarily represent a discrete user, as a single person could manage multiple accounts (e.g. personal and professional accounts). However, we suspect this is unlikely to be the case for most users in the corpus.

6.4.3 Step two: pre-processing the corpus

Having gathered nearly 11 million tweets from known users of te reo Māori, the next step was to clean the data and remove any tweets that would hinder analysis of the RMT Corpus. There were several major sources of noise, including tweets containing little or no Māori text, tweets containing a strong mixture of two or more languages (e.g. English/Māori code-switching), retweets and short tweets lacking sufficient context or meaning. We decided to deal with the non-language-related sources of noise first: those which could be addressed without knowing how much Māori text was in the tweet.

Several adjustments were made to enhance the quality of the corpus. We discarded retweets (re-postings of tweets) because we were not interested in capturing data from *passive* users of te reo, who shared other people's Māori-language tweets without actively writing their own. This also ensured the re-

moval of duplicate tweets that would have negatively skewed the data. 1,522,221 tweets containing fewer than four tokens (words) were also removed, on the basis that they carried minimal linguistic value, even if written completely in Māori (e.g. *Āe!!* meaning "Yes!").

We then applied minor stylistic changes to the formatting of each tweet. In order to improve readability, HTML entities were decoded (e.g. *"* and *<*, representing a quotation mark (") and the less-than sign (<), respectively). These characters were considered to be distracting and unhelpful. The text in the *content_with_emojis* variable preserves special characters (which are useful for tasks such as sentiment analysis), whereas the *content* variable contains only alphanumeric characters, basic punctuation and vowels with macrons.

Adjustments were also made to textual features with special properties, such as *@user* mentions and hyperlinks. In a bid to protect users' identities (following ethics guidelines proposed by Wilkinson and Thelwall, 2011), we replaced all *@user* mentions with "⟨*user*⟩". There are 60,416 ⟨*user*⟩ references in the final corpus, distributed among 40,484 tweets (some tweets mention multiple users). Therefore, just over half of all tweets in the corpus include one or more user mentions. We did not want to completely erase *@user* mentions because they are sometimes integrated into the tweet text itself and can therefore contribute to the syntax/meaning of a tweet. Moreover, when we eventually deployed the rule-based Māori-language classifier (Section 6.4.4), we did not want *@user* mentions to be treated as normal text, because usernames do not inherently have a language, and thus should not influence the percentage of Māori text detected in a tweet. The classifier simply disregards all occurrences of ⟨*user*⟩. In a similar manner, all hyperlinks, beginning with "http://", were standardised with the text "⟨*link*⟩" to denote their position in the tweet. The final corpus contains 29,826 ⟨*link*⟩ references, with 35.74% of all tweets containing one or more links. Links tend to occur at the end of tweets, whereas user mentions tend to occur at the beginning (mainly because replies are automatically prepended with the username/s of the recipient/s). Note that we did not modify hashtags (*#topic*) as these have meaningful linguistic content and may be written in te reo Māori (examples include *#tereo* and *#tekupu*; see Section 6.5.4).

6.4.3.1 Removing specific users

When inspecting the number of tweets per user, it was clear that three religious accounts were dominating the corpus. The purpose of their tweets was to

disseminate existing translations of the Bible; however, they appear to have been controlled by Twitter bots rather than human moderators. All three accounts were suspended by Twitter and have not been active since 2012/2013. In any case, the content of their tweets does not reflect natural, everyday usage of te reo Māori. As such, we removed all tweets posted by these accounts from the RMT Corpus. Similarly, tweets were removed from a fourth account that explicitly mentioned the word "bot". These changes reduced the corpus by 60,011 tweets—an astonishing number considering this amounts to 76% of the final corpus.

6.4.4 Step three: filtering out non-Māori tweets

The third—perhaps most crucial—step was to address the overwhelming presence of non-Māori tweets in our dataset. As noted earlier, most Māori-language tweeters typically only tweet in Māori a fraction of the time, and the majority appear to tweet predominantly in English. There were also many tweets containing both Māori and English text, which is a natural product of bilingual language use, called codeswitching. However, because our goal was to create a high-quality corpus of te reo Māori text, rather than to capture instances of code-switching, we needed a way to filter out these tweets, along with those containing no Māori whatsoever. Our process of selecting tweets is much stricter than the *Indigenous Tweets* website; we filter out roughly three-quarters of the tweets that it classifies as Māori.

Our solution for removing non-Māori tweets in the corpus was to calculate the amount of Māori text in each tweet, and then isolate tweets that fell below a certain threshold, allowing for a small margin of error. In order to do this, it was first necessary to identify and extract the Māori tokens (words) used in each tweet. Existing code developed by Te Hiku Media was adapted for this task. We are grateful to Te Hiku Media for allowing us to use code in the *nga-kupu* repository (Te Hiku Media, 2019), in accordance with the Kaitiakitanga Licence (Te Hiku Media, 2021b).[11]

The *nga-kupu* repository consists of three main Python scripts: *kupu_tūtira*, *hihira_raupapa* and *auaha_tūtira_tū* (see the repository README file for a description of each). The first of these, *kupu_tūtira*, was deemed to be the most relevant for our purposes. *kupu_tūtira* takes as input some corpus text and returns a list of words that are orthographically consistent with Māori. Specifically, a word is considered to be Māori if it meets the following criteria:

[11] *Kaitiakitanga* is a term meaning 'guardianship', as opposed to 'ownership'.

1. The word only contains characters that are part of the Māori alphabet.[12]
2. The word follows consonant/vowel alternation: most Māori syllables take the form (C)(V)V; a few are (C)V1V1V2 (Keegan, 2021).
3. The word does not contain any double consonants, excluding 'ng' and 'wh', which are single consonants in Māori.
4. The word ends in a vowel.

While the *kupu_tūtira* script is effective at identifying Māori words, it does not preclude any non-Māori words that happen to satisfy the above criteria. To resolve this issue, additional items were appended to the stop list, so that they would not be treated as Māori words. These words were determined by manually inspecting the script's output. We coded anything that was obviously English, including a mixture of formal and informal terms: "autotune", "amirite" (am I right?), "epitome", "imitate", "meringue", "nope", "where" and "whiteware" (among others). Our approach relies heavily on having identified Māori-language tweeters in advance (as per step one); otherwise, the algorithm would admit false positives from users who tweet in languages with similar syllable structures.

The *kupu_tūtira* script was then modified to treat each tweet as a separate document/instance. For each tweet, the percentage of Māori text was computed by counting the number of Māori words and dividing this value by the total number of words (see Figures 6.7 and 6.8). All three values (number of Māori words, total number of words and percentage of Māori text) were added to the corpus; this allows analysts to manipulate the data according to the composition of each tweet (e.g. to derive a subcorpus of tweets with a higher threshold, or to analyse tweets of a specific length).

In order to maximise the amount of data available without compromising its quality, we experimented with a range of different thresholds for the percentage of Māori text in each tweet. Following a holistic examination, we opted for two different thresholds based on the length of the tweet: a minimum of 80% Māori text for tweets containing fewer than seven words, and 70% for tweets containing seven words or more (see Equation 6.1). Using these values, we found that 9.4 million tweets did not meet the threshold, and were subsequently removed. This left only 83,491 tweets in the corpus.

$$threshold = \begin{cases} 0.8, & \text{if } w < 7 \\ 0.7, & \text{otherwise} \end{cases} \tag{6.1}$$

[12]The Māori alphabet is derived from the Roman script and consists of five vowels (*a, e, i, o, u*) and ten consonants (*h, k, m, n, ng, p, r, t, w, wh*). A diacritic mark called a macron (tohutō) is placed over vowels to indicate a lengthened vowel sound.

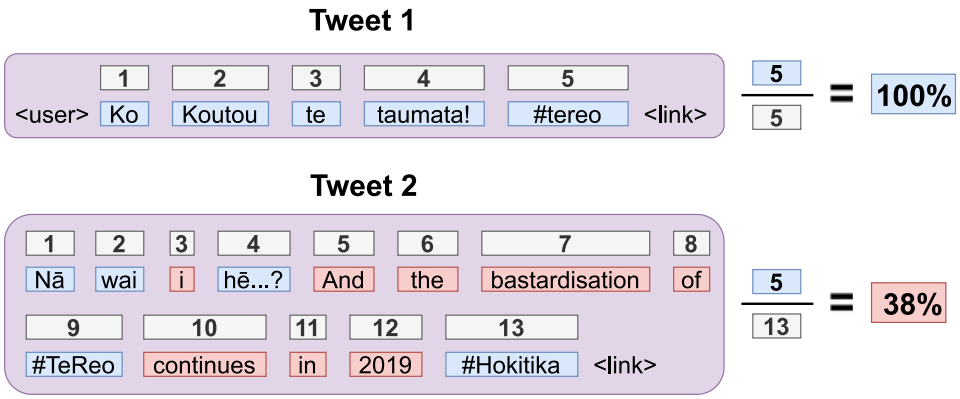


Figure 6.7: Visualisation showing how the percentage of Māori text was calculated for two different tweets. The grey numbered boxes represent the individual words (tokens) that contribute to the final calculation. Blue words are considered to be Māori, whereas red words are considered non-Māori.

To be included in the corpus, tweets are required to have at least 80% Māori text if they contain fewer than seven words (w), and at least 70% Māori text if they contain seven words or more.

We believe these values achieve a sensible trade-off between corpus size and accuracy, with all tweets containing mostly Māori text. A higher threshold, such as 90%, would filter out more tweets containing non-Māori words at the expense of also rejecting a greater number of true Māori tweets. There are valid reasons why a Māori-language tweet might receive a lower score than expected. For instance, a tweet may contain some Māori words that were wrongly classified as non-Māori (see Tweet 2). A tweet might also reference one or more non-Māori entities, better known by their English names, such as a person, place or thing. For example, in the following tweet, "St Pauls" is counted as two non-Māori words: "Te hariru a te Kura o Ōwairaka me te Kāreti o St Pauls mutu ana tā rātou tukinga whiringa whāiti i ngā".

The *kupu_tūtira* script is well-suited to Twitter data for two main reasons. First, the script allows for variations and misspellings, which are a common feature of social media language. For instance, "Aoteroa" is accepted instead of *Aotearoa*, and "whaaaro" is accepted instead of *whaakaro*. These word forms are not unusual on Twitter; however, they are unlikely to be encountered in more traditional genres. Second, there is flexibility concerning double-vowel orthography as an alternative to macron use (e.g. *oo* instead of *ō*). Tweeters sometimes adopt double vowels as a matter of preference,[13] because they may

[13]This convention was popularised by Bruce Biggs in the 1960s, due to his computer's inability to type the macron character (personal communication to Te Taka Keegan, 29 September, 1992).

believe it is a more traditional orthography, or simply because macrons are harder to type. Thus, the script identifies both *Māori* and *Maaori* as words in te reo.

Figure 6.7 illustrates how the percentage of Māori text was calculated for two different tweets, one which was included in the corpus (Tweet 1) and one which was filtered out (Tweet 2). For Tweet 1, all five words were correctly classified as Māori, and the tweet was assigned a score of 100%. Notice that the ⟨*user*⟩ and ⟨*link*⟩ references are excluded from the calculation, which is why they do not have associated indices. This is the case for all tweets, even if the ⟨*user*⟩ reference is syntactically important. The classifier does not attempt to segment hashtags, so *#tereo* is counted as a single word, despite being made up of *te* and *reo*.

Tweet 2 is an example of a mixed-language tweet that was filtered out of the corpus. Five of its thirteen words were deemed to be Māori, yielding a score of 38%. Notice that the third word, *i*, is not counted as a Māori word, when it should be. This tweet also demonstrates that Māori place names are classified as te reo (e.g. *Hokitika*). Conversely, English place names are considered non-Māori, even if they are part of an otherwise completely Māori tweet.

Unsurprisingly, because our chosen thresholds allow 20–30% leeway, there are some tweets that contain a small amount of non-Māori text. Tweet 3 (Figure 6.8) is an example of a tweet that only just meets the 70% threshold. Composed of three English words and eight Māori words, this tweet features a COVID-19 slogan, "Mā tātou katoa e ārai atu te" ("We can all prevent this"). The first word ("wash") is used as a humorous play on "watch". Despite the presence of English, Māori is still the prevailing language in the tweet; an English-only speaker would struggle to understand its meaning. This is true of many other tweets in the corpus that also have relatively low Māori percentages.

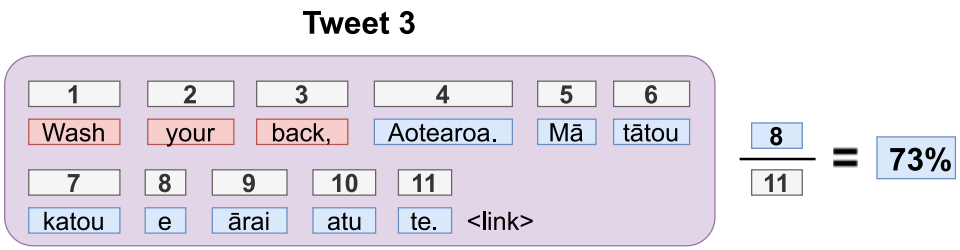


Figure 6.8: A tweet whose percentage of Māori text only just surpassed the 70% threshold.

While the *kupu_tūtira* script was highly effective at removing both non-Māori tweets and tweets containing large quantities of English, it has minor limitations that need to be considered. Most notably, some English words share the same form (spelling) as Māori words, meaning that they are ambiguous cases. Ideally, the classifier would maintain awareness of the surrounding context of these words to decide whether they are Māori (e.g. "hope" is likely to be English if surrounded by English words, but Māori if surrounded by Māori words). However, the classifier treats *all* instances of a word as being either Māori or non-Māori, regardless of actual usage in a particular context.

If a Māori word contains more than three letters and shares the same form as an English word, it is always considered to be Māori. Examples of such words are *hope*, *take*, *more* and *mate*. Although these words tended to be used much more frequently in English, especially because the input dataset contained more English text than Māori, we deduced that these words were more likely to be the difference between a valid Māori tweet being rejected from the corpus than an English tweet being wrongly included.[14]

Conversely, Māori words containing fewer than three letters and sharing the same form as English words were mostly assumed to be non-Māori.[15] This includes the words *i*, *a*, *to* and *no*, which are all very common in English, but which are also common function words in Māori (see Tweet 2). We could have modified the *kupu_tūtira* script so that these were always treated as Māori, but this would have resulted in more English tweets being added to the corpus. This reinforces the need for allowing some flexibility in the percentage of Māori text, because a Māori tweet will invariably be penalised if it contains many instances of these words.

In addition, the *kupu_tūtira* script cannot always distinguish between legitimate Māori words (or their misspellings) and completely made-up words (e.g. *tatata* vs "tatatata"). Fortunately, these made-up words tend to be highly infrequent, which means their impact is negligible. We could have verified the legitimacy of these words using Te Hiku Media's *hihira_raupapa* script, which compares them with the online Māori dictionary (Moorfield, 2021), but we did not want to forego the flexibility that came with permitting common variations and misspellings encountered on Twitter.

[14] An example of a tweet where this is *not* the case is: "Kia ora e hoa, *hope* it's useful!!", where *hope* is regarded as a Māori word, and the tweet receives a final score of 5/7 (71%) instead of 4/7 (57%).

[15] *He* being a notable exception.

6.4.4.1 English versus Māori-lanugage tweeting

Less than 1% of the 9.5 million tweets that were used as input for step three had sufficient Māori text to be included in the corpus. The overwhelming majority of discarded tweets were written in English, showing that the English language dominates the Twittersphere even among the cohort of reo Māori tweeters. Qualitative research would be needed to understand the reasons why Māori-language users choose to tweet in Māori or English in a particular context. For example, an individual's decision to tweet in Māori (or not) might depend on an awareness of their audience, their own proficiency, ideology, the topic of conversation, or a desire to conform with societal norms and expectations. Interviews could be conducted with Māori-language tweeters to gain deeper insight into the rewards and challenges of tweeting in te reo, and how these weigh up against tweeting in English.

6.4.5 Step four: removing formulaic tweets

While significant progress had been made in filtering out unwanted noise, manual inspection of the corpus revealed that some tweets which had been classified as Māori still did not reflect natural usage of te reo. We performed a fourth and final step, removing so-called 'formulaic' tweets, to address this.

When sorting the data alphabetically, we noticed a considerable number of tweets in the corpus whose content was very similar, or even identical. However, this was to be expected, due to the presence of short tweets containing generic phrases and expressions (e.g. "Kia ora e hoa", "Rā whānau ki a koe", "ka mau te wehi"), which are frequently used by both the same and different users. It makes sense to keep these tweets in the corpus, because they reflect authentic language use and may provide insights into the frequency of various syntactic patterns and recurrent constructional schemas.

However, contrary to this, we also encountered a less desirable type of similar tweet, which clearly did *not* reflect natural speech. The content in these tweets appeared to have been created according to some sort of template. Typically, these tweets were identical, except for minor differences in formatting, such as capitalisation, punctuation, numbers, users and links. For example, the tweets "Te Karere Pānui—Rāpare 13 Hakihea 2012—$\langle link \rangle$" and "Te Karere Pānui—Rāpare 06 Hakihea 2012—$\langle link \rangle$" are identical apart from the numbers (dates) after *Rāpare* (13 and 06). Because these tweets can occur a large number of times, their inclusion in the corpus would unfairly inflate associated counts.

To resolve this issue, we transformed tweets into a 'comparison string', and kept only one instance of the (original) text for any longer strings that matched. Each comparison string was generated by converting all letters in the tweet to lower-case, and stripping any punctuation, numbers and ⟨*user*⟩ or ⟨*link*⟩ references. The resulting strings were then compared with all other tweets *by the same user*. We observed that the (undesirable) synthetic tweets tended to be longer than the (desirable) common phrases, which seemed to mostly contain four to six words. As such, we decided to keep all instances of similar tweets containing fewer than seven words, and only the first instance of any similar tweets by the same user containing seven words or more. This also served to remove accidental repostings of longer tweets (e.g. two identical tweets posted within a minute of each other). 3,926 tweets were removed accordingly.

We removed a further 547 formulaic tweets that were not detected during this process, because they only occurred once. This mostly included announcements for events, programs and competitions, such as kapa haka results (e.g. "Kākahu: 3rd Te Rōpū Manaaki, 2nd Te Rōpū Kapa Haka o Whaitara, 1st Te Wharekura o Hoani Waititi #hakawhangarei"). Such tweets contained almost no Māori text other than proper nouns. Tweets were also removed from an account that was dedicated to posting a "Māori word of the day" (e.g. "tapahi(a), tapatapahi(a): tapahi(a), tapatapahi(a): cut, dice. E tapatapahia ana ngā aniana e ia. The... ⟨*link*⟩ #tekupu"). The content in these tweets is not unique to Twitter, being directly copied from (and linked to) another site, much like the Bible translations in Section 6.4.3.1. After removing a total of 4473 tweets in step four, 79,018 tweets remained, which we believe (for the most part) contain original and authentic Māori text.

6.5 Preliminary analysis of the RMT Corpus

In this section, we present a preliminary analysis of the RMT Corpus. We start by providing an overview of the corpus, then analyse the most popular words, topics and hashtags, and finally report the number of tweets and users per year, with particular focus on the ten most prolific tweeters. We intend to build on this analysis by creating visualisations for exploring the RMT Corpus in future work.

A basic overview of the RMT Corpus is given in Table 6.2. The corpus encompasses nearly 80,000 tweets, comprising one million words, written by 2,300 users. Note that hyphenated words are counted as a single token (e.g.

each instance of *here-turi-kōkā* counts as one word). ⟨*link*⟩ and ⟨*user*⟩ references are excluded from the total word count, whereas numbers and hashtags are not.

Table 6.2: Key summary statistics for the RMT Corpus.

Description	**Value**
Number of Tweets	79,018
Number of Tokens (Words)	1,007,652
Number of Users (Tweeters)	2,302
Average Tweet Length (Words)	12.75
Average Tweets per User	34.33
First (Oldest) Tweet	25 May, 2007
Last (Most Recent) Tweet	11 December, 2020
Time Period	13.5 years

6.5.1 Top 100 words by frequency

We now examine high frequency vocabulary in the RMT Corpus. Figure 6.9 is a word cloud showing the 100 most common tokens (words) in the corpus. The size and darkness (hue) of each word are proportional to its overall frequency. When performing these calculations, all tokens were converted to lower-case but otherwise left unchanged. As a result, macron and non-macron variants of the same word (e.g. *māori* and *maori*) were not consolidated, but rather treated as separate entities. An exhaustive word list, with frequencies for all tokens in the RMT Corpus (not just the top 100), is available on the companion website (Kiwi Words, 2021).

Looking at Figure 6.9, it is clear that *te* is by far the most prolific word in the corpus. This is not surprising, given that *te* in Māori has a similar role to the definite article ("the") in English (albeit confined to singular noun phrases): it serves a ubiquitous grammatical function that does not add any meaning to a sentence or clause (Harlow, 2001, p. 309). There is only one hashtag that appears in the top 100 tokens, namely *#tereo*, which is ranked 13th overall, with 10,486 occurrences. 21 words in the top 100 contain one or more macrons, most notably: *ngā* ('the', plural; 7th; 15,294 instances), *māori* ('of Indigenous language/people'; 27th; 5,848 instances), *rā* ('day/sun'/various meanings; 28th; 5,840 instances) and *tēnei* ('this', by speaker; 30th; 5,576 instances).

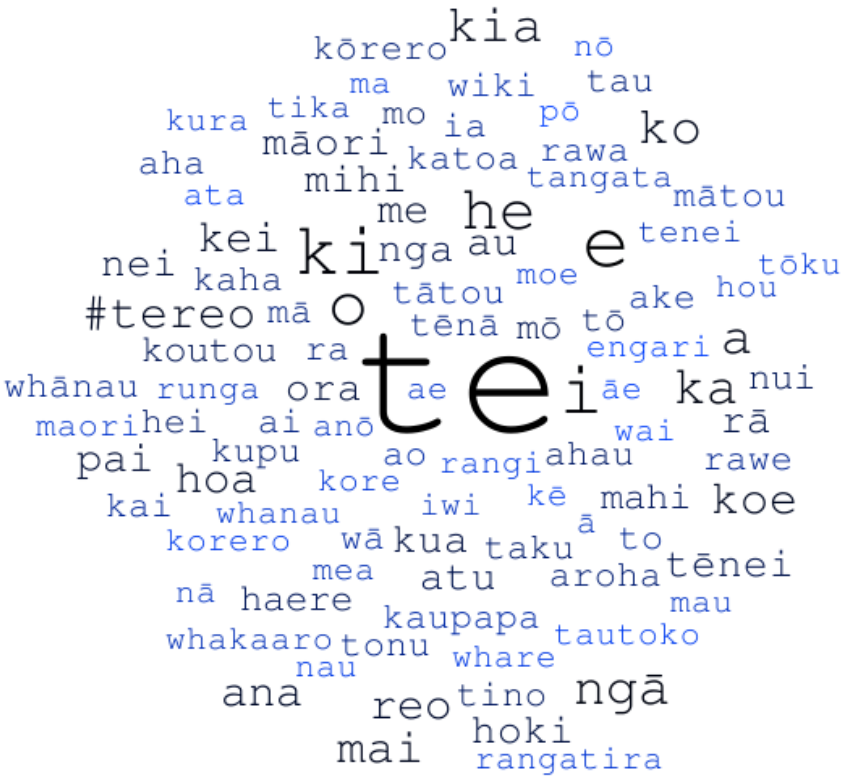


Figure 6.9: Word cloud showing the top 100 words in the RMT Corpus, where size and hue are proportional to frequency. Drawn using `https://worditout.com`.

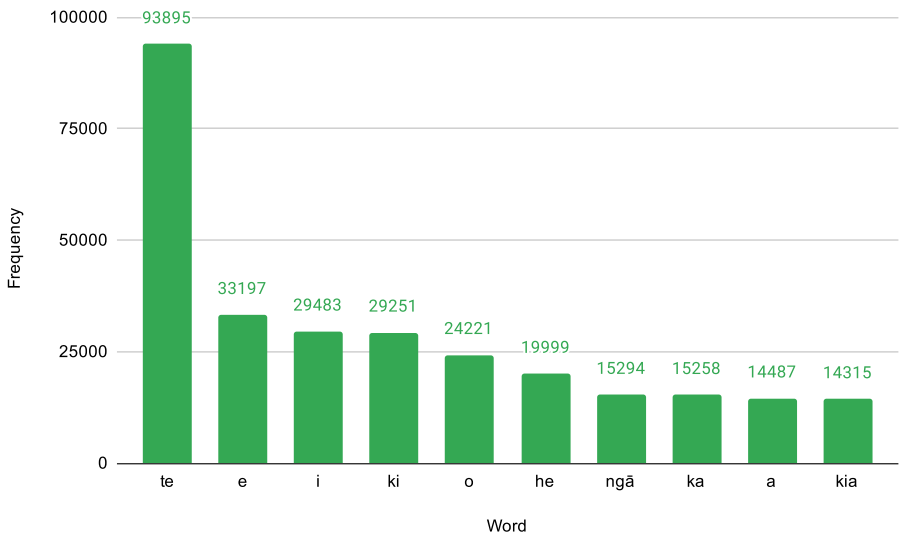


Figure 6.10: The ten most frequent words in the RMT Corpus, which are all function words, and their associated frequencies of occurrence. The most common meanings of these words are: *te* 'the' (singular), *e* (various meanings), *i* (various meanings), *ki* 'to/at', *o* 'of', *he* 'a/some', *ngā* 'the' (plural), *ka* (verbal particle), *a* (nominal particle), and *kia* (various meanings).

6.5.2 Top ten words by frequency

Figure 6.10 drills down into the ten most frequent tokens in the RMT Corpus. As with English (Kilgarriff, 2006), these are all function words (particles) rather than content words. Their role tends to be as determiners, or as words signalling tense, aspect and mood. Like all function words in Māori, they are very short, containing no more than two morae. It is worth noting that the frequency of *i*, *he* and *a* could be conflated with English (arising from tweets with a small number of non-Māori words), although one would still expect these instances to be predominantly Māori.

An interesting hypothesis is that character limits imposed by Twitter might influence the frequency with which these words are used. Boot et al. (2019) found that Dutch-language tweeters adapted the content of their tweets to overcome character limit constraints, changing sentence structure and word forms as necessary. Twitter's character limit increased from 140 to 280 characters in November 2017, so tweets before and after this period may have noticeable linguistic variation (as was investigated in the aforementioned study). Further analysis would be needed to determine whether Māori function words are used differently (e.g. more sparingly) on Twitter, compared to other genres.

6.5.3 Top ten content words by frequency

As the ten most frequent words in the corpus are all function words, we now provide an overview of the top ten content words (Figure 6.11). Many of these words appear to be used in tweets whose discourse function is to increase solidarity, as they are typically used in greetings (e.g. *hoa* 'friend', *pai* 'good', *mihi* 'to greet', *ora* 'healthy', *kaha* 'strong'). Other content words relate more directly to aspects of Māori culture or language (e.g. *māori* 'indigenous', *reo* 'language' and *kupu* 'word'). We excluded the hashtag *#tereo* as this is considered in the next section.

6.5.4 Hashtags in the RMT Corpus

Hashtags are a pervasive feature of Twitter and are widely used in search engines. Users often include hashtags in their tweets to contribute to a wider discussion, introduce new topics, extend conversations or emphasise key points in their tweets. Although introduced to Twitter in August 2007, the first hashtag does not appear in the RMT Corpus until 2009. In total, there are 46,195 hashtag 'tokens' in the corpus, including 10,441 distinct types, although

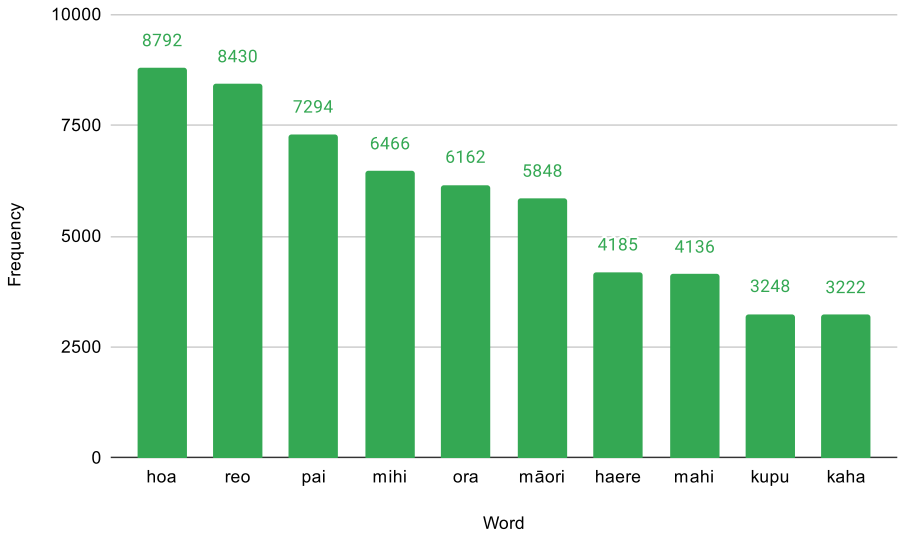


Figure 6.11: The ten most frequent content words in the RMT Corpus and their associated frequencies of occurrence. Common meanings of these words are: *hoa* 'friend', *reo* 'language', *pai* 'good', *mihi* 'to greet', *ora* 'healthy', *māori* 'Indigenous', *haere* 'come', *mahi* 'work', *kupu* 'word' and *kaha* 'strong'.

only 3073 of these appear more than once.[16] These hashtags are written in both Māori and English, and tell us about the general interests and motivations of Māori tweeters.

Figure 6.12 shows the diachronic trajectory of the ten most frequently used hashtags in the RMT Corpus. Semantically, the hashtags pertain mostly to topics concerning Māori language and culture (e.g. *#tewikiotereomaori* refers to Māori Language Week), which was also reported as a prominent theme in relation to hybrid hashtags (Trye et al., 2020). Note that the graph does not include instances of hashtags whose timestamp is unknown (roughly 3% of the data). There are likely to be other tweets (both Māori and non-Māori) that also use these hashtags, but which are not included in the corpus. The data for 2020 is also less reliable than in previous years, for reasons outlined in Section 6.4.2. Consequently, the patterns shown in Figure 6.12 do not necessarily characterise the overall distribution of each hashtag on Twitter. Nevertheless, they provide an indication of when the ten hashtags were most frequently being used, especially with Māori content, and whether they persisted across multi-

[16]Unlike the content and function word analyses above, the number of hashtag types (in this section) was calculated after converting all letters to lower-case and removing macrons. It made sense to merge these different forms, because searching for the hashtag *#Maori* on Twitter will also return results for *#maori* and *#Māori*: `https://twitter.com/search?q=%23Maori&src=typed_query`. Moreover, hashtags did not permit macrons in the early days of Twitter.

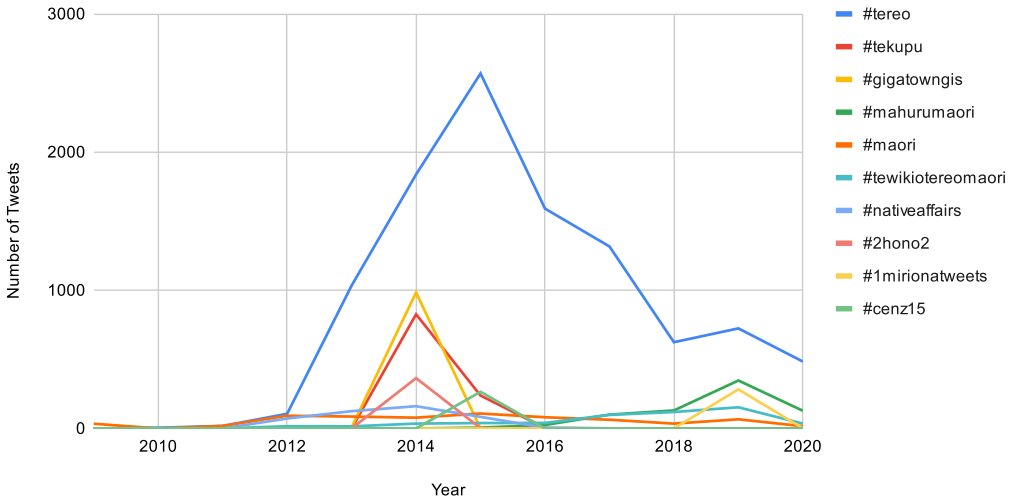


Figure 6.12: Diachronic trajectory of the ten most common hashtags in the RMT Corpus.

ple years. Only half of the top ten hashtags are recorded at least once in the corpus in 2020 (namely, *#tereo*, *#mahurumaori*, *#maori*, *#tewikiotereomaori* and *#1mirionatweets*).

The most dominant hashtag in the corpus is *#tereo* (dark blue), which had more uses than any other hashtag between 2010 and 2020. In 2015 alone, this hashtag appeared in 2,569 tweets. In fact, for any given year between 2014 and 2017 (inclusive), *#tereo* is used more often than the total frequency (i.e. all years combined) for each of the other nine hashtags.

The hashtags *#gigatowngis*, *#tekupu* and *#2hono2* all peaked in 2014. *#gigatowngis* refers to a national competition to receive city-wide funding for ultra-fast broadband, enabling the winner to become New Zealand's first "Gigatown" (Chorus NZ, 2014). *#2hono2* is used in relation to a fortnightly Twitter chat, facilitated by Connected Educator New Zealand (CENZ), with a view to "connecting Māori medium learning environments, whānau, educators, & communities throughout Aotearoa" (CENZ, 2021). This hashtag appears in 364 tweets in 2014, and then only twice in 2015. *#cenz15* refers to the 2015 iteration of the CENZ programme, and, as the name suggests, is used exclusively in 2015.

The hashtags *#mahurumaori* and *#1mirionatweets* both peaked in 2019 (*mahuru* means 'September' and *miriona* means 'million'). This is not surprising, because they were two of the three official hashtags for a nation-wide campaign launched in September 2019, which aimed to achieve one million te reo Māori tweets. Following the launch, Te Māngai Pāho (the Māori Broadcast

Table 6.3: Information about the ten most frequent hashtags in the RMT Corpus, including their frequencies and an example of each.

Rank	Hashtag	Raw Frequency	Distinct Users	Example
1	*#tereo*	10,512	358	Ka tangi te pīpīwharauroa, ko te karere a Mahuru #pēpeha **#tereo** ⟨*link*⟩
2	*#tekupu*	1,168	104	Kia horapa ēnei pānui ēnei ki te rohe o Whanganui **#tekupu** ⟨*link*⟩
3	*#gigatowngis*	1,027	47	Tera, ka huri ki te ara etahi, ko Titirangi i kora **#gigatowngis** ⟨*link*⟩
4	*#mahurumaori*	790	127	#22pressupchallenge, **#mahurumaori**, ra waru ka tukuna hei arohanui ki ku matua kk ka riro atu ki te p. ⟨*link*⟩
5	*#maori*	718	168	Tirohia ngēnei! Te wai korarī nō te rohe o Maniapoto #inu **#Māori** ⟨*link*⟩
6	*#tewikiotereomaori*	597	243	Kia Pai Tō Haere. He rauemi marautanga. **#tewikiotereomāori** ⟨*link*⟩ ⟨*link*⟩
7	*#nativeaffairs*	501	91	**#NativeAffairs** ka aroha ki aua kura.
8	*#2hono2*	374	29	Whoop whoop! Whakarauora te reo māori, kia eke ki tōna panekiretanga **#2hono2**
9	*#1mirionatweets*	289	53	Kia kaha tātau ki te kōrero i te reo Māori, ahakoa ki whea, ahakoa ko wai! **#1MirionaTweets** ⟨*link*⟩
10	*#cenz15*	266	8	Kei runga noa atu koe e hoa! #tereo **#CENZ15** ⟨*link*⟩

Funding Agency) released topics every day during the month of September, which New Zealanders could then weave into their content. It appears that the milestone of one million te reo tweets is still a long way off, with the hashtags *#mahurumaori* and *#1mirionatweets* collectively appearing only 1,079 times in the corpus. The third hashtag for the campaign, *#1mirionatihau*, was ranked 15th overall, with 208 occurrences.

Table 6.3 (on the previous page) provides additional statistics about the top ten hashtags, including their raw frequency (in the corpus) and the number of tweeters who used them at least once. An example of each hashtag is also given to provide additional context about how it is used. *#tereo* is not only used in the most tweets, but also by the most distinct users (358). *#tewikiotereomaori* is ranked sixth overall, but is used by the second highest number of users (243). The hashtag *#cenz15* is only employed by eight different tweeters, but is used more frequently among them, with an average of 33 instances per tweeter.

6.5.5 Basic user statistics

Different users have different numbers of tweets in the RMT Corpus, depending on how often they tweet in te reo. A diachronic analysis of the ten most active tweeters is given in Section 6.5.7, but we first examine the overall composition of the corpus at a macro-user level. Figure 6.13 shows that the distribution of tweets by Māori-language users is highly skewed, with a small proportion of users contributing many tweets to the corpus, and a large proportion having only one or a few. Thus, a small number of Twitter users account for the

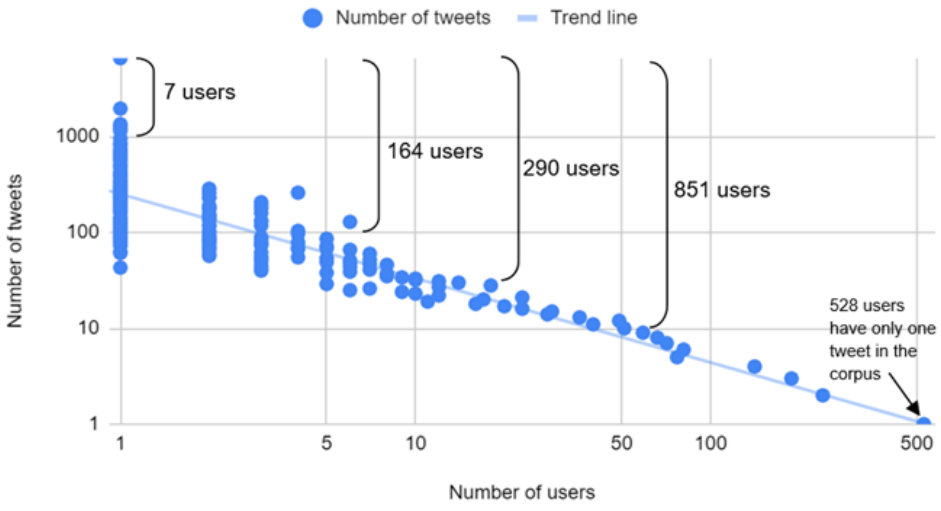


Figure 6.13: The number of tweets per user in the RMT Corpus, shown on a logarithmic scale and revealing a highly skewed distribution.

majority of Māori-language tweets, which also matches the distribution on the *Indigenous Tweets* website. Specifically, the top eight users account for a fifth of the data in the RMT Corpus; the top 24 users account for a third; and the top 70 account for half.

Figure 6.14 plots the number of new Māori-language tweeters per year in the RMT Corpus. In this graph, tweeters are only counted the year that their very first tweet appears in the corpus. This provides an indication of the extent to which the community of Māori-language tweeters has grown over time, although there is no guarantee that tweeters from past years are still active, and some users may have deactivated previous accounts and created new ones, thereby distorting the figures. Overall, we can see that there is some degree of positive growth across all years (i.e. there is at least one new Māori-language tweeter per year), but that the period from 2012 to 2018 experienced the most growth, with more than 200 new users per year. 2012 had the largest surge compared to the previous year. The uptake of new users in later years has been comparatively slow, with only 42 new users recorded for 2020 (also representing the largest *decrease* with respect to the previous year; however, this could also be an artefact of the *Indigenous Tweets* classifier not having been updated recently). There are 140 tweeters for whom no tweet timestamps are available, whose data is therefore not shown.

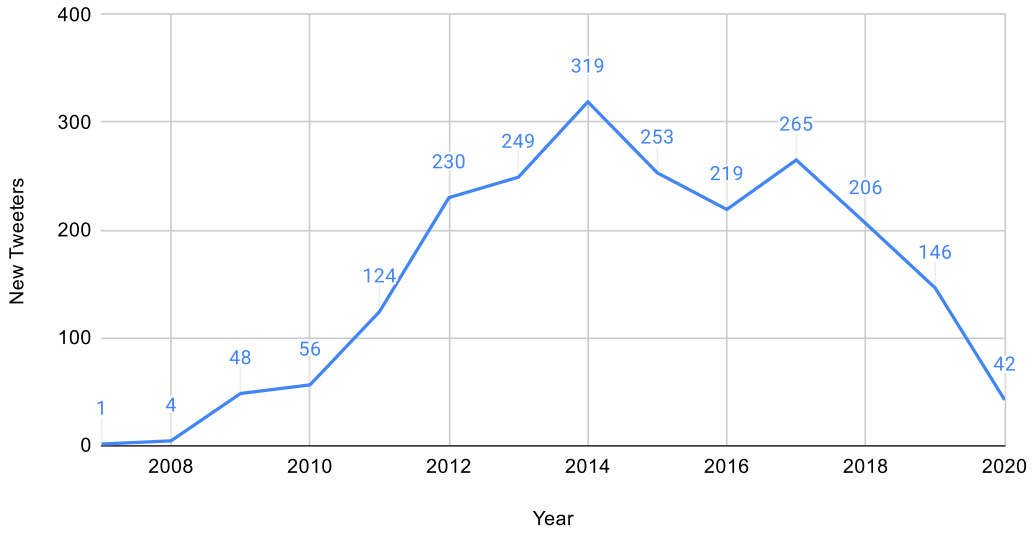


Figure 6.14: The number of *new* users per year in the RMT Corpus, representing growth in the Māori-language Twitter community over time.

6.5.6 Diachronic analysis of tweets per year

Next, we discuss how the number of Māori-language tweets and active users in the RMT Corpus changes over time, and comment on related patterns and trends. Figure 6.15 shows the distribution of tweets per year, whereas Figure 6.16 shows the active users. The number of tweets and active users appear to be positively correlated (i.e. as the number of users increases, so does the number of tweets in the corpus).

It is immediately apparent from Figure 6.15 that very few Māori tweets included in the corpus were posted during the first few years of Twitter (to be precise: none in 2006, one in 2007, 9 in 2008 and 245 in 2009). The volume of Māori tweets increases very steeply over the next five years, reaching its highest point, 10,622 tweets, in 2014. Following this, the number of tweets declines for a couple of years, and then remains relatively stable, with reasonably small fluctuations from year to year. By 2020, although the number of tweets is roughly only two-thirds of what it was in 2014, there are still significantly more tweets than there were pre-2013, a trend which looks likely to continue.

It should be noted that 6,115 tweets (7.74% of the corpus) are omitted from Figure 6.15 because we were not able to retrospectively retrieve their timestamps. Therefore, a limitation of these findings is that they may not reflect actual trends for the entire corpus, especially if a large portion of excluded tweets were posted in the same year. In addition, a limitation of the data for 2020 is that, due to changes in how the tweets were collected, we could not gather tweets from some users whose data was previously available. This means that the figure reported for 2020 is likely to be an underestimate of the actual amount, and we must therefore proceed with caution.

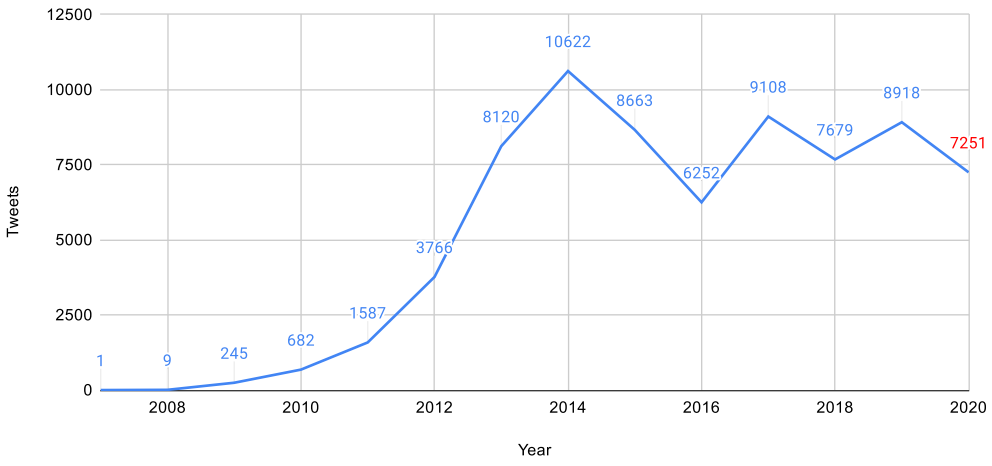


Figure 6.15: The number of tweets per year in the RMT Corpus.

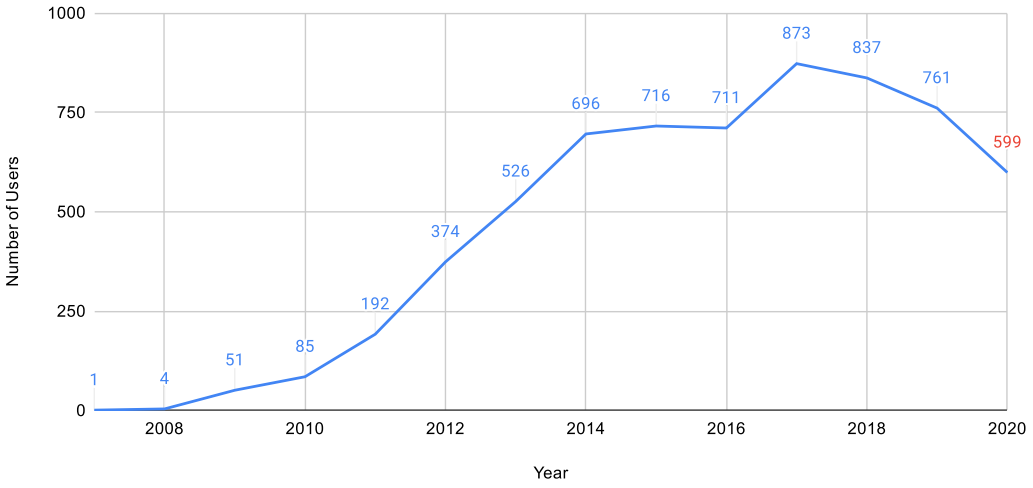


Figure 6.16: The number of active users per year in the RMT Corpus.

Figure 6.16 shows that, even though the number of new users in 2020 is relatively low (Figure 6.14), there is still a large number of active tweeters (599, which is a conservative measure, given the less reliable data for 2020). Comparing Figures 6.15 and 6.16, the number of tweets and active users both increase rapidly between 2009 and 2014. However, unlike the number of tweets, the number of active users in 2016 only decreases very slightly. Thus, it would appear that, in 2016, roughly as many users were tweeting in Māori (compared to 2015), but they were posting fewer tweets each. The number of active users declines steadily over the last four years, from 873 users in 2017 to 599 in 2020.

6.5.7 Diachronic analysis of the ten most active tweeters

Having considered the overall number of tweets and users per year, we now analyse diachronic trends of the ten most prolific tweeters in the RMT Corpus (Figure 6.17). One of the most striking observations is that many of these users have not been as active in recent years. Most of the ten most prolific tweeters' Māori-language activity spiked between 2013 and 2015, after which their volume of tweets steadily declined. While the reasons for this are not clear, it may be because the individuals concerned are now using Twitter less often (perhaps favouring other social media platforms), and as such, posting fewer tweets. However, it is worth noting that most of these users continue to tweet in te reo Māori (in 2020), albeit to a lesser extent. The prevalence of tweets between 2013 and 2015 matches and, indeed, goes some way towards explaining, the overall temporal trends seen in Figure 6.15.

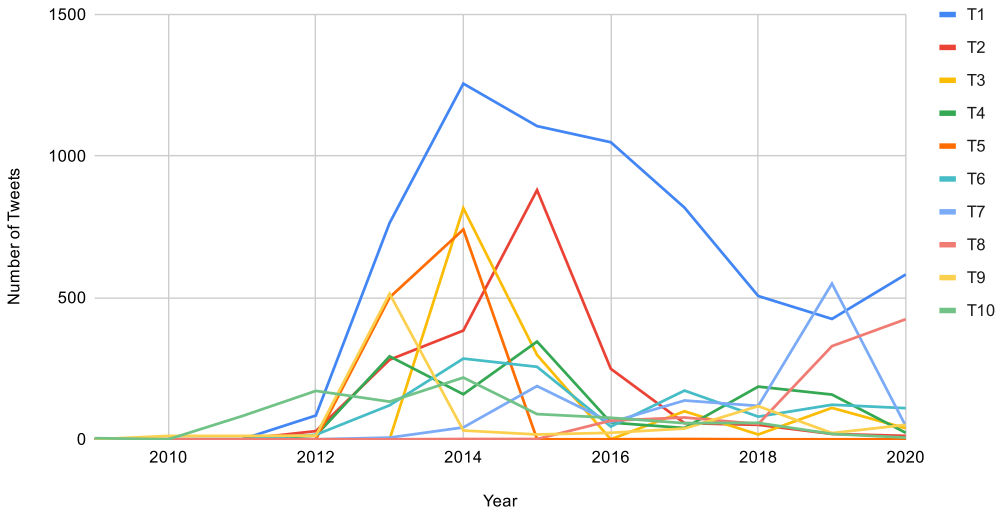


Figure 6.17: Diachronic trajectory of the ten most frequent tweeters in the RMT Corpus.

6.6 Downloading the RMT Corpus

The RMT Corpus can be downloaded for academic use. Please follow the instructions, given in both Māori and English, on our companion website: `https://kiwiwords.cms.waikato.ac.nz/rmt_corpus/`. In order to comply with Twitter's terms of service, we have only released the IDs and selected metadata for each tweet. Some tweets in the RMT Corpus are no longer publicly available and, as such, cannot be downloaded from Twitter. The statistics and analyses presented in this paper reflect version one of the corpus (rmt-corpus-v1.csv); however, we intend to supplement our data with new, more recent tweets in the future. See the project GitHub repository for further details: `https://github.com/Waikato/kiwiwords/tree/master/rmt_corpus`.

6.7 Conclusions and future work

The main contribution of this work is a publicly available corpus of reo Māori tweets, which constitutes the largest known collection of Māori-language data from social media. The RMT Corpus was compiled by gathering tweets from users identified by the *Indigenous Tweets* classifier. Although 11 million tweets were initially collected, only 79,018 of these (0.72%) satisfied our stringent criteria for inclusion in the corpus. Tweets of many different types, including non-Māori tweets, short tweets, retweets, automated tweets and formulaic tweets, were all filtered out at various stages, in order to improve corpus relia-

bility. The final corpus comprises roughly one million words, written by 2302 users.

Our preliminary analysis examined high-frequency words and hashtags in the RMT Corpus. Initial findings suggest that Māori-language tweets tend to reference topics relating to Māori language and culture. These tweets are often used to increase solidarity and signal community affiliation. We also investigated the number of tweets and active users per year, paying special attention to the ten most prolific tweeters, who account for more than a fifth of the data in the corpus.

The RMT Corpus opens exciting avenues for future work. It demonstrates how te reo Māori is used in authentic, communicative contexts, which makes it valuable for linguistic analysis and the creation of teaching materials. The corpus constitutes a rich, multi-dimensional dataset, whose user metadata could be used to inform decisions about language revitalisation initiatives, especially in digital contexts. From a linguistics perspective, the RMT Corpus can be used (among other things) to study high frequency words, collocations and constructional schemas.

From a sociolinguistic viewpoint, it would be interesting to investigate whether the corpus captures any recent changes in the Māori language, such as a shift in the use of pronouns and possessive markers (A/O categories) or voicing distinctions. Given that Māori is experiencing a wave of changes, whereby L1 *kaumātua* (elders) are using different constructions from L2 young Māori-language learners, it would be interesting to investigate how these changes might be evolving in a platform which largely appeals to a younger audience.

As a unique source of social media data, the RMT Corpus could be fruitfully compared with other, more traditional genres of Māori text (including those detailed in Section 6.3), for instance using a keyword analysis (following Hardie, 2014). Analysts could also study inter-speaker variation within the corpus, both synchronically and diachronically, by drawing on the rich array of tweet and user metadata (although care would need to be taken to derive a balanced subset for analysis).

Qualitative research is needed to determine the reasons why Māori-language speakers choose to tweet in Māori (or not) in a particular context, and to shed light on the rewards and challenges of doing so. Furthermore, we anticipate that the RMT Corpus will facilitate the development of new NLP resources for the Māori language, including text-to-speech, speech-to-text and auto-completion technologies.

As has been emphasised throughout the paper, the RMT Corpus was de-

signed to contain almost exclusively Māori text. An empirical evaluation of our language identification scheme is left for future work. Our next step will be to create a corpus of Māori/English code-switching from the discarded tweets (among others), by targeting those with mixed-language text. We believe this will have useful applications in NLP, such as helping to improve New Zealand English text-to-speech, by capturing how Māori is spoken in everyday contexts (i.e. often interspersed with English).

Finally, the RMT Corpus could benefit from extra data, perhaps even collecting this in real-time. Tweets posted after December 2020 would make a welcome addition to the corpus, as would Māori-language data from other social media platforms, such as Facebook and Instagram. We currently discard bilingual tweets, even if they contain Māori text that could stand alone, so it might be beneficial to incorporate this data in the future. Having as much high-quality language data as possible is vital to ensure the accuracy and subsequent proliferation of te reo Māori in NLP applications and environments.

Acknowledgements

This research was made possible by funding from Ngā Pae o Te Māramatanga and the University of Waikato. We thank Tamahau Brown for his valued contribution to a pilot study. We are indebted to Kevin Scannell, who generously gave his time and resources to support this research, especially with matters concerning *Indigenous Tweets*. Te Hiku Media kindly granted us permission to reuse existing code, which proved instrumental in cleaning the corpus. We thank Andreea Calude, Jonathan Dunn, David Bainbridge and three anonymous reviewers for their insightful comments and suggestions, which helped improve the paper. Any remaining errors are our own.

6.8 Postscript

Motivated by the scarcity of resources available for the Māori language, this chapter has introduced a corpus of (primarily) monolingual tweets, called the RMT Corpus, which contains rich metadata and presents new opportunities for linguistic analysis. Consequently, as mentioned earlier, this dataset will form the basis of a quantitative study of linguistic possession in Māori (Chapter 8). Before we come to this, however, we will explain how we created additional resources by leveraging the mixed Māori-English tweets that were set aside during the cleaning process described in this chapter.

6.9 References

Apperley, M., Keegan, T. T., Cunningham, S. J., and Witten, I. H. (2002). Delivering the Maori-language newspapers on the internet. In Curnow, J., Hopa, N., and McRae, J., editors, *Rere atu, taku manu! Discovering history, language & politics in the Maori-language newspapers*, pages 211–232. Auckland University Press.

Bender, E. (2019). The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Bender, E. M., Hovy, D., and Schofield, A. (2020). Integrating ethics into the NLP curriculum. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 6–9.

Bible Society New Zealand (2021). Te Paipera Tapu - Māori Bible. `https://biblesociety.org.nz/discover-the-bible/the-bible-in-maori/maori-bible-app/`.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Bloem, J., Versloot, A., and Weerman, F. (2019). Modeling a historical variety of a low-resource language: Language contact effects in the verbal cluster of early-modern Frisian. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 265–271.

Boot, A. B., Sang, E. T. K., Dijkstra, K., and Zwaan, R. A. (2019). How character limit affects language usage in tweets. *Palgrave Communications*, 5(1):1–13.

Boyce, M. (2011). Mana aha? exploring the use of mana in the legal Māori corpus. *Victoria University of Wellington Law Review*, 42(2):221–240.

Boyce, M. and Stephens, M. (2016). The legal Māori corpus: Texts printed before 1910. `https://nzetc.victoria.ac.nz/tm/scholarly/tei-legalMaoriCorpus.html`.

Boyce, M. and Stephens, M. (2022). The legal Māori corpus. `https://www.legalmaori.net/corpus`.

Boyce, M. T. (2006). *A Corpus of Modern Spoken Māori*. PhD thesis. Unpublished PhD thesis available in the library at Victoria University of Wellington.

Cassels, M. (2019). Indigenous languages in new media: Opportunities and challenges for language revitalization. *Working Papers of the Linguistics Circle*, 29(1):25–43.

CENZ (2021). Connected Educator New

Zealand. http://connectededucator.org.nz/event/twitter-chat-2hono2-te-tuhonohono-i-te-iwi-maori/.

Chorus NZ (2014). #gigatown. https://company.chorus.co.nz/gigatown.

Cocks, J. and Keegan, T. T. (2014). The Māori macron restoration service. http://community.nzdl.org/macron-restoration/jsp/en/main.jsp.

Coto Solano, R., Nicholas, S., and Wray, S. (2018). Development of natural language processing tools for Cook Islands Māori. In *Proceedings of Australasian Language Technology Association Workshop*, pages 26–33.

Cunliffe, D., Morris, D., and Prys, C. (2013). Investigating the differential use of welsh in young speakers' social networks: A comparison of communication in face-to-face settings, in electronic texts and on social networking sites. In Jones, E. H. G. and Uribe-Jongbloed, E., editors, *Social media and minority languages: Convergence and the creative industries*, pages 75–86. Multilingual Matters.

El-Haj, M., Kruschwitz, U., and Fox, C. (2015). Creating language resources for under-resourced languages: Methodologies, and experiments with arabic. *Language Resources and Evaluation*, 49(3):549–580.

Finn, A. (2021). Whakairo kupu: Te reo Māori part-of-speech tagger. In *Māori Speech Hui*, University of Auckland, Auckland.

Giles, H., Bourhis, R. Y., and Taylor, D. M. (1977). Towards a theory of language in ethnic group relations. In Giles, H., editor, *Language, ethnicity and intergroup relations*, pages 307–348. Academic Press.

Google (2006). Google Translate. https://translate.google.com/.

Grey, S. G. (1928). *Ngā mahi a ngā tūpuna*. 3 edition.

Hardie, A. (2014). Log ratio–an informal introduction. Corpus approaches to social science. http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/.

Harlow, R. (2001). *A Māori reference grammar*. Longman.

Harlow, R. (2007). *Māori: A linguistic introduction*. Cambridge University Press.

Harlow, R. B. and Barbour, J. (2013). Māori in the 21st century: Climate change for a minority language? In Vandenbussche, W., Jahr, E. H., and Trudgill, P., editors, *Language Ecology for the 21st Century: Linguistic Conflicts and Social Environments*, pages 241–266. Novus Press.

Innes, F. (2021). Online typing assistance for te reo Māori. Unpublished honours dissertation at the University of Waikato.

James, J., Shields, I., Berriman, R., Keegan, P. J., and Watson, C. I. (2020). Developing resources for te reo Māori text to speech synthesis system. In

International Conference on Text, Speech, and Dialogue, pages 294–302. Springer, Cham.

Jones, D. B., Robertson, P., and Taborda, A. (2015). Corpus of Welsh language tweets. http://techiaith.org/corpora/twitter/?lang=en.

Ka‘ai, T. (2017). Te whare matihiko o te reo-digital tools for the revitalisation of te reo māori. In Whaanga, H., Keegan, T. T., and Apperley, M., editors, *He whare hangarau Māori-language, culture & technology*, pages 29–34. Te Pua Wānanga ki te Ao/Faculty of Māori and Indigenous Studies, Te Whare Wānanga o Waikato/University of Waikato.

Ka‘ai, T., Ó Laoire, M., and Ostler, N. (2012). Language endangerment in the contemporary world: Globalisation, technology and new media. In Ka‘ai, T., O Laoire, M., Ostler, N., Ka‘aiMahuta, R., Mahuta, D., and Smith, T., editors, *Language Endangerment in the 21st Century: Globalisation, Technology and New Media*, pages 1–4. Foundation for Endangered Languages & Te Ipukarea - The National Maori Language Institute, AUT University, Auckland, New Zealand.

Keegan, P. J. (2021). Māori language information. https://www.maorilanguage.info/mao_vocab_faq.html.

Keegan, T. T., Hudson, M., and Mahelona, K. (2021). Data sovereignty. In *Language and Society Conference 2020, University of Waikato*. https://www.youtube.com/watch?v=sOps3_tEXGE&list=PLp619EeWvHk7OQkGqVsfhcoB744wsSDLb&index=6.

Keegan, T. T., Mato, P., and Ruru, S. (2015). Using Twitter in an indigenous language: An analysis of te reo Māori tweets. *AlterNative*, 11(1):59–75.

Keegan, T. T. A. G. and Cunliffe, D. (2014). Young people, technology and the future of te reo Māori. In Higgins, R., Rewi, P., and Olsen-Reeder, V., editors, *The value of the Māori language: Te Hua o te Reo Māori*, pages 385–398. Huia Publishers.

Kilgarriff, A. (2006). BNC database and word frequency lists. https://www.kilgarriff.co.uk/BNClists/lemma.num.

King, B. P. (2015). *Practical natural language processing for low-resource languages*. PhD thesis, University of Michigan.

King, J. (2018). Māori: revitalization of an endangered language. In Rehg, K. L. and Campbell, L., editors, *The Oxford handbook of endangered languages*, pages 592–612. Oxford University Press.

King, J., Maclagan, M., Harlow, R., Keegan, P., and Watson, C. (2010). The MAONZE corpus: Establishing a corpus of Maori speech. *New Zealand Studies in Applied Linguistics*, 16(2):1–16.

Kiwi Words (2021). Putunga reo Māori tīhau: The reo Māori twitter corpus. https://kiwiwords.cms.waikato.ac.nz/rmt_corpus/.

Kukutai, T. and Taylor, J. (2016). *Indigenous data sovereignty: Toward an agenda*. ANU Press.

Lynn, T. and Scannell, K. (2019). Code-switching in Irish tweets: A preliminary analysis. In *Proceedings of the Celtic Language Technology Workshop*, pages 32–40.

Mato, P. and Keegan, T. T. (2013). Indigenous tweeting for language survival: The Māori-language profile. *International Journal of Technology and Inclusive Education*, 2(2):184–191.

Maxwell, M. and Hughes, B. (2006). Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the COLING/ACL 2006 Workshop on Frontiers in Linguistically Annotated Corpora*. Association for Computational Linguistics.

May, S. and Hill, R. (2018). Language revitalization in Aotearoa/New Zealand. In *The Routledge Handbook of Language Revitalization*, pages 309–319. Routledge.

McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., and McCullough, D. (2012). On building a reusable Twitter corpus. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1113–1114.

Mereraiha, H. (2019). 1mil tweets in te reo. https://www.teaomaori.news/1mil-tweets-te-reo.

Meyerhoff, M. (2019). *Introducing sociolinguistics*. Routledge.

Moorfield, J. C. (2021). Te aka Māori dictionary. https://maoridictionary.co.nz/.

Moses, C., Thompson, M., Mahelona, K., and Jones, P.-L. (2020). Scoring pronunciation accuracy via close introspection of a speech recognition recurrent neural network. Poster session at NeurIPS 2020. https://papareo.nz/docs/PapaReo_NeurIPS2020_Poster.pdf.

NZDL (2002). Welcome to the Māori Niupepa collection. https://www.nzdl.org/cgi-bin/library.cgi?a=p&p=about&c=niupepa.

Scannell, K. (2011a). Indigenous tweets. http://indigenoustweets.com/.

Scannell, K. (2011b). Welcome/fáilte! http://indigenoustweets.blogspot.com/2011/03/welcomefailte.html.

Scannell, K. P. (2007). The crúbadán project: Corpus building for under-resourced languages. In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, volume 4, pages 5–15.

Scannell, K. P. (2022). Managing data from social media: The Indigenous Tweets project. In Berez-Kroeker, A. L., McDonnell, B., Koller, E., and Collister, L. B., editors, *The Open Handbook of Linguistic Data Management*. MIT Press.

Sciascia, A. D. (2016). Māori cultural revitalisation in social networking sites. Paper prepared for Te Puni Kōkiri.

Shields, I., Watson, C., Keegan, P., Berriman, R., and James, J. (2019). Creating a synthetic te reo Māori voice. In *International Conference on Language Technology for All*, Paris.

Stats NZ (2020). 2018 census totals by topic – national highlights (updated). https://www.stats.govt.nz/information-releases/2018-census-totals-by-topic-national-highlights-updated.

Te Hiku Media (2019). Identify Māori text. https://github.com/TeHikuMedia/nga-kupu.

Te Hiku Media (2021a). Corpus of te reo derived from the New Zealand Hansard. https://github.com/TeHikuMedia/nga-tautohetohe-reo.

Te Hiku Media (2021b). Kaitiakitanga — guardianship. https://papareo.nz/#kaitiakitanga.

Te Hiku Media (2021c). Natural language processing tools for te reo Māori. https://papareo.io/.

Te Taura Whiri i te reo Ma¨ori (2012). Guidelines for Māori language orthography.

Trye, D., Calude, A., Bravo-Marquez, F., and Keegan, T. T. (2019). Māori loanwords: A corpus of New Zealand English tweets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 136–142, Florence. Association for Computational Linguistics.

Trye, D., Calude, A. S., Bravo-Marquez, F., and Keegan, T. T. (2020). Hybrid hashtags: #youknowyoureakiwiwhen your tweet contains Māori and English. *Frontiers in Artificial Intelligence*, 3:15.

Twitter (2021a). Conversation id. https://developer.twitter.com/en/docs/twitter-api/conversation-id.

Twitter (2021b). How to get the blue checkmark on Twitter. https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts.

Twitter (2021c). Twitter API. https://developer.twitter.com/en/docs/twitter-api.

Twitter (2021d). The Twitter rules. https://help.twitter.com/en/

rules-and-policies/twitter-rules.

Verhoeven, B., Daelemans, W., and Plank, B. (2016). Twisty: A multilingual twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1632–1637.

Waitangi Tribunal (1986). Report of the Waitangi Tribunal on the te reo Māori claim. Retrieved from https://forms.justice.govt.nz/search/Documents/WT/wt_DOC_68482156/Report%20on%20the%20Te%20Reo%20Maori%20Claim%20W.pdf.

Waitoa, J., Scheyvens, R., and Warren, T. R. (2015). E-whanaungatanga: The role of social media in māori political empowerment. *AlterNative*, 11(1):45–58.

Whaanga, H. (2020). Ai: a new (r)evolution or the new colonizer for indigenous peoples. In Lewis, J., editor, *Position paper on Indigenous Protocol and Artificial Intelligence*, pages 34–38. The Initiative for Indigenous Futures and the Canadian Institute for Advanced Research (CIFAR).

Wilkinson, D. and Thelwall, M. (2011). Researching personal information on the public web: Methods and ethics. *Social science computer review*, 29(4):387–401.

Zaghouani, W. and Charfi, A. (2018). Arap-tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Zuckermann, G. A. (2020). *Revivalistics: From the genesis of Israeli to language reclamation in Australia and beyond*. Oxford University Press.

Chapter 7

A Hybrid Architecture for Labelling Bilingual Māori-English Tweets

This chapter introduces linguistic resources tailored to the unique bicultural context of Aotearoa New Zealand. By drawing on tweets that were filtered out of the RMT Corpus (Chapter 6), we develop a hybrid architecture for classifying the source language of individual words and full sentences in bilingual Māori-English texts. This architecture integrates machine learning models with hand-crafted rules based on orthographic features of Māori. Following extensive fine-tuning, we deploy the architecture to create an annotated corpus of mixed Māori-English tweets, which we call the MET Corpus. We also develop custom interactive visualisations for exploring the MET Corpus and analysing the different kinds of classification errors, which involve several categorical variables.

Publication Details

The following paper has been reproduced with minor changes to the formatting, as discussed in Section 1.4:

Trye, D., Yogarajan, V., König, J., Keegan, T. T., Bainbridge, D., & Apperley, M. (2022, November). A hybrid architecture for labelling bilingual Māori-English tweets. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022* (pp. 119-130). https://aclanthology.org/2022.findings-aacl.11

Abstract

Most large-scale language detection tools perform poorly at identifying Māori text. Moreover, rule-based and machine learning-based techniques devised specifically for the Māori-English language pair struggle with interlingual homographs. We develop a hybrid architecture that couples Māori-language orthography with machine learning models in order to annotate mixed Māori-English text. This architecture is used to label a new bilingual Twitter corpus at both the token (word) and tweet (sentence) levels. We use the collected tweets to show that the hybrid approach outperforms existing systems with respect to language detection of interlingual homographs and overall accuracy. We also evaluate its performance on out-of-domain data. Two interactive visualisations are provided for exploring the Twitter corpus and comparing errors across the new and existing techniques. The architecture code and visualisations are available online, and the corpus is available on request.

7.1 Introduction

"Ko te reo te mauri o te mana Māori.
Ko te kupu te mauri o reo Māori."

Translated to English as *The language is the life force of the mana Māori. The word is the life force of the language* (Higgins and Keane, 2015), this famous saying by Tā Hēmi Hēnare (Sir James Hēnare) encapsulates the importance of the Māori language to Māori, the Indigenous people of Aotearoa[1] New Zealand.

Te reo Māori is both endangered and low-resourced, with limited corpora and Natural Language Processing (NLP) techniques available (James et al., 2020). Data annotation currently has to be done manually by language experts, making the process time-consuming and resource-intensive. These obstacles hinder technological advances that could assist in maintaining the language and, consequently, the culture of Māori.

The Māori language used today is frequently interspersed with English, either in the form of *code-switching* (Holmes and Wilson, 2017; Marras Tate and Rapatahana, 2022) or *borrowing*. Here, the borrowing process is bidirectional, resulting in both English loanwords in Māori (Harlow, 1993) and Māori loanwords in English (Calude et al., 2020). The latter are not only used by bilingual Māori speakers, but also by monolingual English-speaking New

[1] Aotearoa is increasingly used as a Māori name for New Zealand. Te reo Māori means 'the Māori language'.

Zealanders. Linguists are interested in determining the frequency of these patterns, which are reflective of Aotearoa New Zealand's unique bicultural identity.

The interweaving of Māori and English is a key consideration for developing robust technologies that can accommodate practical, everyday usage of te reo Māori and New Zealand English. Leveraging the abundance of relevant data on Twitter, our research focuses on the following task:

> *Automatic language identification for bilingual Māori-English text at both the token (word) and tweet (sentence) level.*

Differentiating between Māori and English text is not straightforward. This is because both languages use the Roman script, and *interlingual homographs*—words that are spelt the same but differ in meaning across languages (Dijkstra, 2007)—are prolific. These words present a major challenge for classifying mixed-language text, especially if they are highly frequent in both target languages (Barman et al., 2014). Consider the following tweets in which interlingual homographs are emphasised:

(a) **Here** is **to a more** productive day tomorrow
(b) Ka **kite** koe **i a** koe!
(c) **He** is at **a** tangi in Ruatoki. Doubt **he** did

In terms of annotation, the desired tweet-level labels are (a) English, (b) Māori, and (c) Bilingual. These are determined with recourse to the individual token labels: all tokens in (a) are English, all tokens in (b) are Māori, and (c) contains a mixture of tokens from both languages, with 'tangi' (funeral) and 'Ruatoki' (a place name) being labelled Māori. According to our approach, all words of Māori origin are tagged as Māori, even if they are used as borrowings in English.

In order to obtain accurate tweet and token-level labels, we utilise knowledge and understanding gained from Māori researchers, Māori technology developers and the Māori community. Our methodology involves combining machine learning techniques with Māori orthography, thereby instantiating the pipeline recommended by Hämäläinen (2021). We hypothesise that doing so will improve the overall accuracy of language identification for bilingual Māori-English text.

This paper makes the following contributions:

1. Development of a hybrid architecture (Trye et al., 2022a) to detect Māori and English words for a given bilingual text input.
2. The *Māori-English Twitter (MET) Corpus*, a first-of-a-kind dataset comprising bilingual and monolingual tweets, annotated at the token- and

tweet-level by deploying our architecture.

3. Evidence that the hybrid architecture improves both language detection of interlingual homographs and overall accuracy when compared with two existing techniques.
4. Two interactive visualisation tools for exploring the corpus and comparing label errors across the different systems.

7.2 Background and Related Work

7.2.1 Māori Data Sovereignty

The Māori language is the natural medium through which Māori express their cultural identity, construct the Māori worldview and convey their authenticity (Marras Tate and Rapatahana, 2022; Rapatahana, 2017; White, 2016). It is crucial to highlight that Māori data needs to be handled with care, because of the injustices caused by colonisation and its effect on the vitality of the language (Smith, 2021). We strongly believe that any NLP resources that are developed from this research, either directly or indirectly, should be created for the good of the Māori-language community and not for the capital gain of others; more generally, Indigenous data should not be commodified at the expense of Indigenous communities (Bird, 2020).

7.2.2 Challenges and Bias in Māori NLP

Key challenges in developing Māori speech and language technology arise from the lack and limitations of resources (James et al., 2020), phonological differences from English, and the lexical overlap between written Māori and English, including more than 100 interlingual homographs (Te Hiku Media, 2022). These obstacles hinder NLP advances that could facilitate the maintenance of Māori language and culture.

Existing large-scale technologies such as cloud-based language-detection tools and voice assistants are predominantly designed for English. These tools fail to recognise or correctly pronounce Māori words, even when used as borrowings in New Zealand English (James et al., 2022b). Our goal is to redress that inequity in NLP resources, and thus mitigate the bias that existing tools have towards the more dominant English language.

7.2.3 Code-Switching in NLP

Bilingual and multilingual code-switching, especially between resource-rich and low-resourced languages, has gained traction as a challenging but important NLP problem (Aguilar et al., 2020; Molina et al., 2016; Solorio et al., 2014). A myriad of studies investigating code-switching on social media has emerged, showcasing challenges and possibilities for many different language pairs (Jose et al., 2020; Maharjan et al., 2015; Barman et al., 2014).

While an overview of Māori-language corpora is given in Trye et al. (2022), we detail three that are particularly relevant here. The *Hansard Dataset* (James et al., 2022a) comprises two million Māori, English and bilingual sentences, annotated by hand at both the token and sentence levels. The *MLT Corpus* (Trye et al., 2019) is a publicly-available collection of English tweets with Māori borrowings, albeit lacking token-level labels. The *RMT Corpus* (Trye et al., 2022) contains predominantly-Māori tweets and is also publicly-available. We use the hand-crafted rules from the RMT Corpus to detect candidate Māori words based on Māori orthography (Section 7.3.2).

Research using machine learning techniques for te reo Māori is relatively young, and is restricted by the limited scope of available resources. Although cloud-based services offered by corporations such as Google and Microsoft support Māori-language detection, the accuracy of these services is poor (Keegan, 2017; James et al., 2022b).

Recently-developed language identification and code-switching detection models for the Māori-English pair make use of Skipgram-based fastText models to pre-train embeddings (Dunn and Nijhof, 2022; James et al., 2022b). James et al. combine pre-trained embeddings with recurrent neural networks (RNNs) to identify Māori text and code-switching points between the Māori-English pair. Their embeddings were pre-trained on a large collection of bilingual and monolingual data, and shown to outperform open-sourced English-only equivalents. Our hybrid architecture uses the fastText pre-trained embeddings and Hansard training set from James et al. (2022b).

7.3 Methodology

This section details the process used to collect Twitter data (Section 7.3.1) and the techniques underpinning our hybrid architecture. We combine language rules (Section 7.3.2) with neural networks (Section 7.3.3), as suggested by Hämäläinen (2021).

7.3.1 Data Collection and Pre-processing

In order to create a bilingual Twitter corpus on which to deploy our architecture, we combined tweets that were originally gathered for the RMT Corpus with more recent tweets from the same users.[2] Tweets that included 30-80% Māori text under the RMT system were chosen, as it was deduced these would primarily contain instances of Māori-English code-switching. The collected tweets were pre-processed to mitigate noise in the dataset. A series of tweets was removed, including retweets, similar and identical tweets, tweets posted by bots, and tweets containing fewer than four words. Non-Roman characters were stripped from the remaining tweets and common English contractions were expanded. 20,000 foreign-language tweets were then removed via manual and automatic checks, which included searching for symbols denoting glottal stops in the middle of tokens (characteristic of several Polynesian languages related to, but distinct from, Māori). This yielded 178,192 tweets in total. Finally, when extracting the tokens in each tweet, links, user mentions, hashtags, punctuation, emoticons and Arabic numerals were all ignored. The rationale for excluding hashtags is that they often contain abbreviations and/or multiple words, sometimes even combining languages (Trye et al., 2020), making them difficult to annotate without additional pre-processing.

7.3.2 Hand-Crafted Rules

Trye et al. (2022) employ hand-crafted rules to identify Māori tokens in tweets, referred to as the *RMT system* throughout this paper. This technique adapts hand-crafted rules implemented by Te Hiku Media (2019), an Indigenous Māori organisation. The rules are as follows:

- Tokens must contain only characters from the Māori alphabet, which comprises five vowels (*i, e, a, o, u*) and ten consonants (*p, t, k, m, n, ng, wh, r, w, h*).
- Lengthened vowels may be indicated with a macron ($\bar{a}$), or using double-vowel orthography (*aa*).
- Tokens must adhere to Māori syllable structure: they must follow consonant/vowel alternation, end with a vowel, and be free of consonant clusters (excluding the digraphs *ng* and *wh*).
- For input to the algorithm, some further adjustments were made to identify as many candidate Māori words as possible.[3]

[2] Users were identified via *Indigenous Tweets* (Scannell, 2011).

[3] Words like 'a', 'i', 'to' and 'no' were omitted from the original RMT system due to their high frequency in English.

When applied to bilingual text, a major limitation of these rules is that tokens of the same type are always classified the same way (typically as Māori), which is problematic for interlingual homographs.

7.3.3 Machine Learning Component

The hybrid architecture uses Bidirectional Gated Recurrent Units (Cho et al., 2014) with an attention layer as the machine learning component. Text is represented using fastText (Bojanowski et al., 2017) Skipgram-model word embeddings (Mikolov et al., 2013) with 300 dimensions, pre-trained on a collection of Māori and bilingual corpora (James et al., 2022b). The attention layer used is based on the Bahdanau attention mechanism (Bahdanau et al., 2015). Our preliminary experiments favoured the use of Bi-GRU with an attention layer over other deep learning models such as CNNs and LSTMs.

To the best of our knowledge, there is no large bilingual Twitter dataset annotated accurately by experts at the token- or tweet-level. Hence, for training Bi-GRU, we use the Hansard Dataset containing transcribed formal Māori and English (James et al., 2022b). The Bi-GRU model is trained to predict Māori, English or bilingual sentences, using default settings in Keras/Tensorflow. Adam (Kingma and Ba, 2015), an adaptive learning rate optimisation algorithm, was employed as the optimiser for the networks. Softmax activation is leveraged in the output layer. To avoid over-fitting, we use a combination of dropout (Srivastava et al., 2014) with a rate of 0.5 and early stopping (Zhang et al., 2017).[4]

7.4 Hybrid Architecture

The hybrid architecture for labelling bilingual Māori-English datasets at both the token (word) and tweet (sentence) levels builds upon the RMT and ML techniques described in the previous section. Figure 7.1 outlines the process used to label the tweets in our cleaned dataset, and references the algorithms in Appendix E. The architecture can also be directly applied to Māori-English corpora with longer text sequences (Trye et al., 2022a).

7.4.1 Token-Level Labels

Multiple techniques are used to determine the appropriate label for each token (Algorithms 1 and 2). Initially, tokens are deemed to be Māori only if they

[4]Model trained on 12 core Intel(R) Xeon(R) W-2133 CPU @ 3.60GHz, GPU device GV100GL.

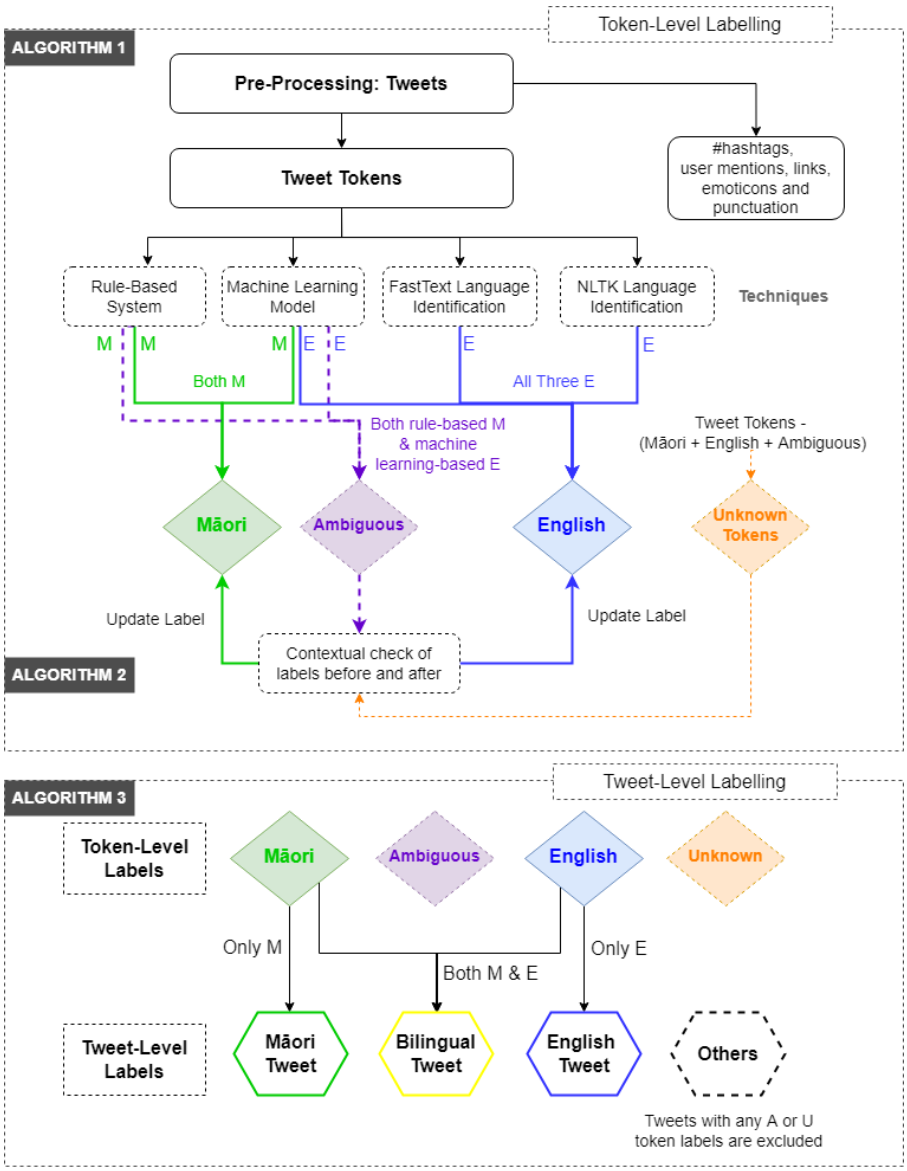


Figure 7.1: Flow chart detailing token- and tweet-level labelling.

are labelled 'M' by both the modified rules from the RMT Corpus and the pre-trained machine learning model. In a similar vein, English tokens are labelled by combining the outcome of using the machine learning model with fastText (Joulin et al., 2017, 2016) and NLTK (Bird and Loper, 2004) language identification models. These techniques have proven high accuracy in detecting English, providing confidence in the 'E' labels. Due to the informal nature of tweets, the language-specific tags include colloquial language and textspeak (e.g. 'u' for 'you' in English).

Any tokens that are labelled 'M' by the modified RMT system and 'E' by the machine learning model are initially classified as ambiguous. The knowledge gained from neighbouring tokens is then used to re-classify these words as Māori or English (Algorithm 2). Crucially, the MET Corpus only includes

tweets comprising 'M' and 'E' token-level labels; all remaining tokens that could not be re-classified with certainty led to the removal of the corresponding tweet, and are left for future research.

7.4.2 Tweet-Level Labels

The updated token labels are used to generate appropriate tweet-level labels (Figure 7.1, Algorithm 3). If a tweet consists solely of 'M' or 'E' tokens, then the tweet-level label is Māori or English, respectively. Tweets that contain at least one 'M' and 'E' token are considered bilingual; this includes single-word borrowings in otherwise monolingual contexts. For further confidence, the tweet-level labels were compared with the pre-trained machine learning model, and it was found that 90% of these labels matched the hybrid model.

7.5 The Māori-English Twitter Corpus

The steps detailed in the previous two sections resulted in the formation of a new bilingual dataset: the *Māori-English Twitter (MET) Corpus*. Key summary statistics for this collection of 76,000 tweets are presented in Table 7.1. Almost 90% of tweets in the corpus are labelled Bilingual, 10% are English and only 0.1% are Māori. This distribution is expected, given the chosen threshold and characteristics of the RMT system used to filter tweets in the data collection phase. In terms of individual words, 60% of tokens in the MET Corpus are labelled English and 40% are Māori. The 20 most frequent tokens are shown in Figure 7.2. Most of these tokens are function words rather than content words, apart from 'Māori' and 'reo' (language), whose presence would suggest that many tweets in the corpus pertain specifically to Māori language and culture.

Table 7.1: Summary statistics for the *MET Corpus*.

	Tweets	**Bilingual (B)**	**English (E)**	**Māori (M)**
Tweets	76,416	67,713	7847	856
Tokens	781,381	-	465,292	316,089
Users	2417	2347	1148	283
Avg tokens/tweet	10	11	6	6
Avg tweets/user	32	29	7	3

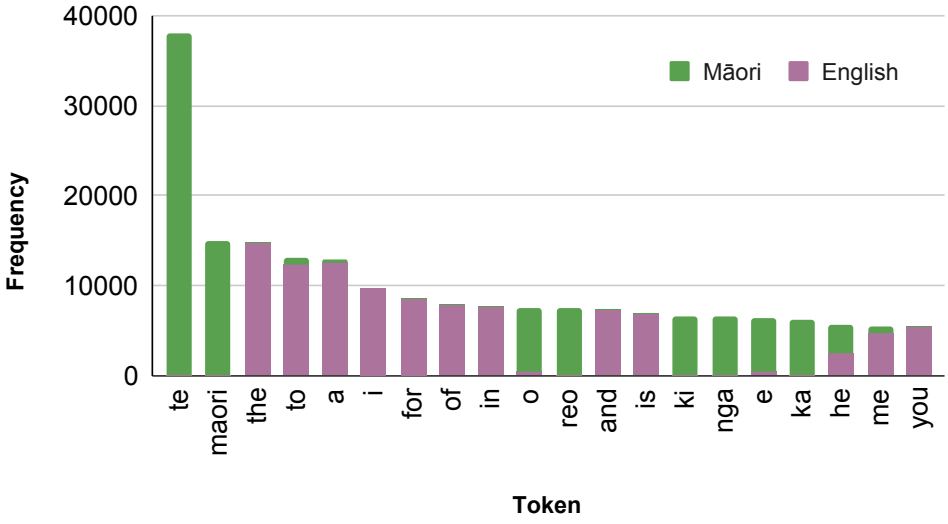


Figure 7.2: The 20 most frequent tokens in the *MET Corpus*: **Māori words**, **English words** and homographs (mixed).

7.5.1 Visualisation of the MET Corpus

We provide an interactive visualisation for exploring the MET Corpus (Trye et al., 2022c); see Figure 7.3. The visualisation includes a scrollable table of tweets and allows the user to select and filter data according to several dimensions. Key features include a treemap (and associated search bar) displaying token frequencies for the selection, a line chart of the distribution of selected tweets over time, and a bubble chart summarising the relative contribution of each user. In addition, selections can be made on both the tweet and token-level labels. The percentage of tweets that is currently visible (with respect to the entire corpus) is indicated at the top left of the display.

7.5.2 Gold Standard Labels

A manual annotation process was used to obtain gold standard labels for a random one percent sample of the data (N=850 tweets), including tweets that were ultimately filtered out of the corpus. This process consisted of two phases. In phase one, two of the authors manually tagged the true tweet-level label of each tweet in the sample, so that this could be compared against the predicted label for each system. Furthermore, the coders identified which tokens, if any, had been mislabelled by each system. Tokens were considered to be Māori if they were listed in the Māori dictionary (Moorfield, 2021), constituted Māori slang (e.g. 'ktk' is the Māori equivalent of 'lol'), or were Māori named entities. It was decided that even Māori borrowings in otherwise English tweets should

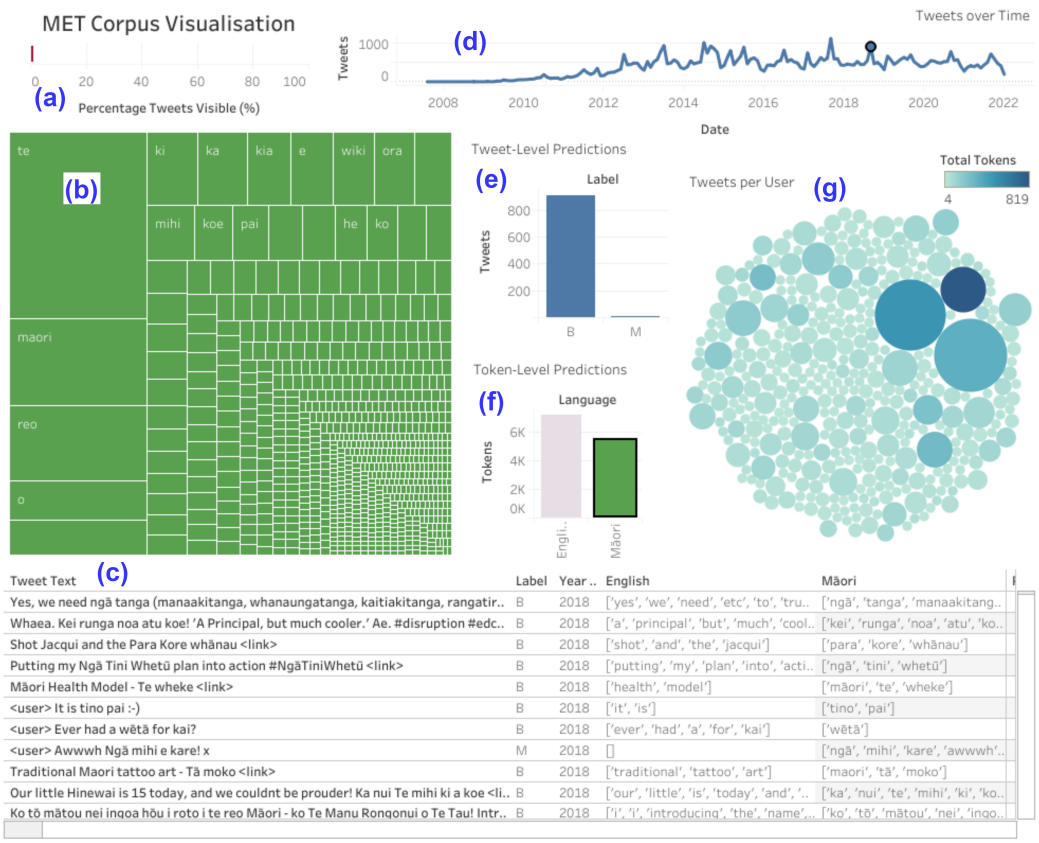


Figure 7.3: Interactive tool for exploring the *MET Corpus*: (a) percentage of corpus visible, (b) selected tokens by frequency, (c) tweet table, (d) tweets by year, (e) tweet predictions, (f) token predictions, (g) tweets by user. Created using Tableau.

be tagged as Māori, because applications such as a New Zealand English text-to-speech tool would be required to correctly identify and pronounce words of Māori origin, regardless of how they are categorised from a theoretical point of view.

In the sample tweets, the coders encountered five foreign tweets (0.6%), which were discarded, since the individual tokens could not be accurately tagged as either English or Māori. In order to assess the efficacy of phase one of the annotation process, Cohen's kappa was computed for a subsample of 200 tweets. This yielded a score of 0.816, indicating a strong level of agreement.

For the second phase, one of the authors went through the data again, and, for each mistaken token, noted whether it was a Māori token that had been mislabelled as English (false negative), or an English token that had been mislabelled as Māori (false positive). Where possible, they recorded further information about the specific type of error. Common error types

included short-length homographs, named entities (including names of people, places, tribes, organisations and events), the presence of one or more non-Māori characters, misspellings and missing macrons.

7.6 Experiment Results and Analysis

This section compares the performance of the newly-developed hybrid system with the standalone RMT (Trye et al., 2022) and ML (James et al., 2022b) systems. We also use a test set from the Hansard Dataset (James et al., 2022a) to evaluate our hybrid architecture with data from another domain. For brevity, we refer to interlingual homographs simply as *homographs*.

7.6.1 Visualisation of System Errors

To facilitate analysis of our manually-coded sample of tweets (hereafter, the *Twitter sample*), we have developed an interactive tool for comparing errors between the three systems of interest (Trye et al., 2022b). The visualisation helps users to explore the relationship between the tweet- and token-level labels for each system, and to better understand which kinds of tokens are responsible for the errors. Figure 7.4 provides a screenshot of this interactive tool, which guided the subsequent analysis.

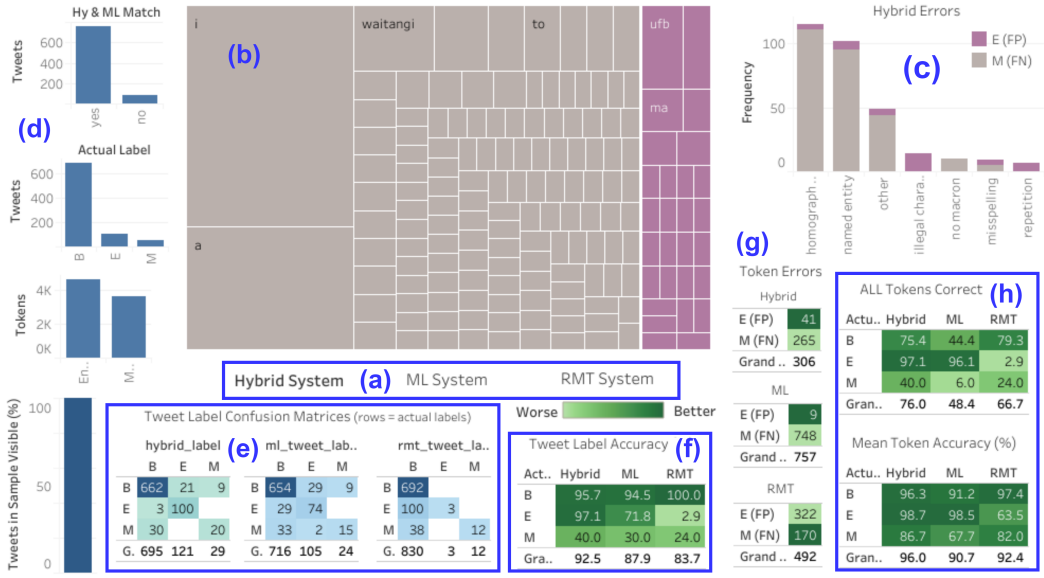


Figure 7.4: Interactive tool for comparing system errors: (a) navigation menu, (b) misclassified tokens, (c) error types, (d) filtering by labels, (e) tweet label confusion matrices, (f) tweet accuracy, (g) token mistakes, (h) token accuracy. Created using Tableau.

7.6.2 Overall Accuracy

Table 7.2 characterises the state of play for the hybrid system and the two existing systems, using six example tweets. All token-level errors are given, together with the resulting tweet labels. The token-level errors obtained using the RMT system's hand-crafted rules are mostly homographs, whereas those for the ML system are mostly Māori words. The hybrid architecture performs well by comparison, correctly identifying all but one Māori token.

Table 7.2: Example tweets indicating **actual Māori tokens**, **tweet-level errors** and unidentified Māori tokens.

	Tweet Labels				Token-Level Errors (FP, FN)		
Tweets	**Actual**	**RMT**	**ML**	**Hybrid**	**RMT**	**ML**	**Hybrid**
1. Teaching ate me alive ⟨*link*⟩ via ⟨*user*⟩ #classroom-reality	E	B	E	E	ate, me	-	-
2. ⟨*user*⟩ **ka pai**! Some **reo** and hugs! What more does one need:) #BFC630NZ	B	B	B	B	more, one	-	-
3. ⟨*user*⟩ ⟨*user*⟩ **Kia ora** Bronwyn. Hope to catch up while we are here!	B	B	B	B	hope, here	Kia	-
4. ⟨*user*⟩ **Ata marie** John, hope you're well mate.	B	B	B	B	marie, hope, mate	-	
5. **E hoa ma, nga mihi o te tau hou!** #Matariki #MaoriNewYear #BNZatm #respect ⟨*link*⟩	M	M	B	M	-	E, o, tau	-
6. **Maori** Party welcomes **Waitangi** Tribunal report	B	B	B	B	-	Waitangi	Waitangi

Table 7.3 provides a synopsis of the system evaluations, broken down by tweet/sentence and token labels for both the Twitter sample and the Hansard test set. Looking at the Twitter sample, the Hybrid system has the highest overall accuracy. The Hybrid system's F1-scores are consistently better than the other two systems' at both the tweet and token level. The specificity of the Hybrid system is good across all tweet-level labels. Notably, the RMT system's specificity is extremely poor for bilingual tweets, indicating that the system is overly eager to find a positive result, even when it is not present. All systems do poorly at identifying Māori-only tweets; most are classified as Bilingual instead. This is likely because 'i' and 'a' are frequent in Māori but nearly always classified as English.

The Hansard test set included 10,000 bilingual, 1,000 Māori and 1,000 English sentences. The sentence-level accuracy for the RMT system is much better than the other systems; this is likewise true of the F1-scores for both Māori and bilingual sentences. This is because the test set contains predominantly bilingual sentences, and in most cases the RMT system identifies at least one Māori and English token. However, the Hybrid system still has superior specificity for bilingual sentences. At the token-level, the Hybrid system does best for English tokens and the RMT system does best for Māori tokens.

Table 7.3: Tweet and token-level system evaluation for both the Twitter sample and Hansard test set. Recall (R), precision (P), F-score (F1), specificity (S) and overall accuracy are presented, with **best scores** emphasised.

System	Twitter Sample																	
	Tweet-Level													Token-Level				
	English				Māori				Bilingual				Overall	English			Māori	
	F1	P	R	S	F1	P	R	S	F1	P	R	S	Accuracy	F1	P	R	F1	P	R
RMT	0.06	**1.00**	0.03	**1.00**	0.39	1.00	0.24	**1.00**	0.91	0.83	**1.00**	0.10	0.84	0.90	0.93	0.87	0.87	0.88	0.85
ML	0.71	0.70	0.72	0.97	0.40	0.62	0.30	0.98	0.93	0.91	0.95	0.60	0.88	0.94	**0.94**	0.94	0.85	**0.96**	0.79
Hybrid	**0.89**	0.83	**0.97**	0.96	**0.51**	0.69	**0.40**	0.98	**0.95**	**0.95**	0.96	**0.78**	**0.93**	**0.95**	**0.94**	**0.95**	**0.94**	0.92	**0.97**
	Hansard Test Set																	
	Sentence-Level													Token-Level				
RMT	0.33	**0.71**	0.21	0.88	**0.96**	**1.00**	**0.91**	**1.00**	**0.95**	**0.91**	**0.95**	0.55	**0.92**	0.87	0.91	0.84	**0.86**	**0.86**	**0.86**
ML	**0.60**	0.43	0.97	**0.91**	0.32	**1.00**	0.19	0.99	0.79	0.90	0.70	0.55	0.68	0.92	0.91	0.91	0.66	0.70	0.64
Hybrid	0.52	0.35	**1.00**	0.89	0.38	**1.00**	0.24	0.99	0.85	**0.91**	0.79	**0.64**	0.77	**0.93**	**0.92**	**0.92**	0.71	0.73	0.70

7.6.3 Error Analysis

Figure 7.5 and Table 7.4 present a summary of token-level errors in the Twitter sample for all three systems, and highlight errors specifically caused by homographs. All systems struggle with short-length homographs (comprising fewer than five letters) like 'i' and 'a', which are pervasive in both languages. Nevertheless, the hybrid system fares considerably better than the other systems, with the ML and RMT systems having nearly double and over triple the number of homograph errors, respectively.

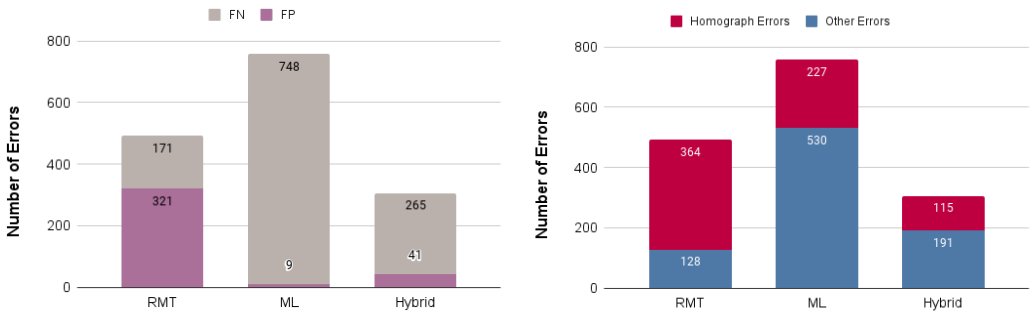


Figure 7.5: Token-level errors in the Twitter sample, showing **false positives**, **false negatives** and **homograph errors**.

The vast majority of errors in the Hybrid system are Māori words that are mislabelled as English. Among these false negatives, short-length homographs constitute 42% of mistakes and named entities constitute 35%. While these are the two largest groups of errors, the Hybrid system still consistently classifies many of these kinds of words correctly (e.g. 'hope', 'Aotearoa').

Table 7.4: Common token-level errors in the Twitter sample, including **homographs**.

System	False Positives	False Negatives
RMT	me, one, more, he, make, here, hope, take, o, nana, u	i, a, to, marie, no, ō, noho
ML	nana, ma	o, e, kia, i, he, a, tau, makaurau, waitangi, me, tūhoe, waatea, au, mo, kai, ō, to, kohanga, matatini, no, ā, morena, horipū, tuhoe
Hybrid	nana, ma, ufb	i, a, waitangi, waatea, to, no, tau, tuhoe

Table 7.5: Common token-level errors in the Hansard test set, including **homographs** mislabelled as 'M'.

System	Hansard Token-Level Errors
RMT	we, are, he, one, more, where, take, here, make, too, rate, none, rape, hope, reiterate, moe, mai, oki
ML	death, moe, mai, rā, hiamoe, kui, ki, te, pō, oti, atu, ai
Hybrid	moe, mai, rā, kui, ki, te, pō, oti, atu, ai

These results indicate that the errors produced by the Hybrid system occur on a smaller scale than the ML system and are easier to fix than those for the RMT system. For instance, it is straightforward to update the labels for all tokens that contain non-Māori characters (like 'ufb'), and named entity accuracy (for tokens such as 'Waitangi') could be improved using an exhaustive list of non-ambiguous Māori place names.

A breakdown of the most prolific errors in the Hansard test set is given in Table 7.5. The most commonly misclassified homographs in both corpora are 'i', 'a', 'to' and 'no', which are all Māori particles that tend to be classified as English. Typically, such words are embedded inside larger segments of Māori text, so it is surprising that these instances are not correctly identified by our hybrid system's contextual check. One of the potential reasons is because the ML component of our hybrid architecture always classifies these tokens as English.

Like the Hybrid system, the ML system tends to mislabel Māori words as English rather than English words as Māori. Many of the same kinds of errors occur, though there are more false negatives and fewer false positives. The ML system frequently misclassified the particles 'e', 'o' and 'kia' in phrases such as "Miharo **e** hoa!", "Te Wiki **o** Te Reo Maori" and "**kia** ora". In contrast, the Hybrid system always labelled these correctly.

The RMT system differs from the others in that it has more false positives than false negatives. As a rule-based system, it always assigns the same label to each word type, even if it is valid in both languages. Words that are consistent with Māori orthography are generally tagged as Māori; as a result, the RMT system is considerably better at correctly classifying Māori named entities, including personal and place names. However, the RMT system performs considerably worse than the other two when classifying tweets with a large proportion of English text. Over 85% of false positives are short-length homographs, with 'me', 'one', 'more', 'he', 'make' and 'here' being the worst offenders. Like the other two systems, there are also some instances of Māori words that are misclassified as English (especially 'i', 'a', and 'to'), due to the stoplist that was used.

7.7 Limitations

The research presented in this paper has some limitations that need to be acknowledged. The hybrid architecture uses a single neural network-based model, but we have experimented with variations in the neural networks and parameter choices. Given the available data and resources, bidirectional RNNs performed the best.

We found that our hybrid architecture does not label Māori named entities consistently, and short-length homographs like 'i' and 'a' are problematic. This requires further investigation, perhaps involving a special look-up for Māori place names, and ensuring that a context check is always carried out for frequent homographs, especially function words.

In addition, our approach for identifying foreign-language tweets is not exhaustive, and in some cases, tokens that are neither Māori nor English will have been erroneously labelled as such. Our foreign-language processing currently focuses on manually identifying problematic tweets in a small subset of the data, then extrapolating this into the wider dataset. This approach could be further developed, or a more automated system could be implemented.

Our labels do not distinguish between borrowings and code-switches (Alvarez-

Mellado and Lignos, 2022). This means it is not possible to automatically extract tweets where Māori borrowings are used in otherwise English contexts, or vice versa, although the number of tokens identified in each language could serve as a useful proxy.

Finally, we discarded a proportion of the collected tweets as our algorithm was not optimised for dealing with undue levels of noise. The discarded tweets with unknown labels are not vital to the MET Corpus presented in this research; however, they require further investigation, and may constitute useful additions to the corpus.

7.8 Conclusions and Future Work

This paper presents an architecture for labelling bilingual Māori-English text, by bringing together machine learning and knowledge of Māori orthography, an approach that could also be fruitful for other endangered languages. We use this architecture to create the first large-scale corpus of bilingual Māori-English tweets annotated at both the token and tweet level. Both this corpus and the Hansard Dataset are used to illustrate the strengths of our approach, including superior token-level accuracy, especially with respect to interlingual homographs. In particular, the specificity scores for bilingual data favour the Hybrid system, while highlighting a major weakness of the RMT system. Additional insights can be gleaned from two exploratory visualisations for interrogating the corpus and comparing system errors.

Future work towards enhancing the bilingual corpus could involve extending this research to classify hashtags as these are currently ignored. Moreover, the architecture lends itself to annotating other bilingual datasets, such as the MLT Corpus (Trye et al., 2019), and could assist in the creation of new resources. A further avenue of exploration would be assigning part-of-speech tags to each token in the corpus, based on the language identified. This could be achieved using newly-developed tools for Māori (Finn et al., 2022) in conjunction with established part-of-speech taggers for English. Such developments are important for ensuring better representation of the Māori language in digital applications and environments.

Acknowledgements

We are indebted to researchers from the University of Auckland and Te Hiku Media for kindly sharing the Hansard Dataset. We thank Andreea Calude and three anonymous reviewers for their helpful comments and suggestions. DT acknowledges funding from the University of Waikato Doctoral Scholarship.

7.9 Postscript

To summarise, this chapter has introduced a hybrid architecture for classifying bilingual Māori-English text that surpasses known alternatives, together with an annotated dataset of mixed-language tweets, which was a natural extension of the RMT Corpus. The corpus provides evidence of the multifaceted ways in which Māori words are used in day-to-day conversation, interwoven with English as single-word borrowings and multi-word code-switches, according to the linguistic knowledge of both the speaker and their intended audience. Although we do not utilise the MET Corpus in subsequent chapters, it presents a number of opportunities for analysing code-switching and language-mixing in Aotearoa New Zealand and ensuring these are better represented in NLP technologies.

7.10 References

Aguilar, G., Kar, S., and Solorio, T. (2020). Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1803–1813.

Alvarez-Mellado, E. and Lignos, C. (2022). Borrowing or codeswitching? Annotating for finer-grained distinctions in language mixing.

Bahdanau, D., Cho, K. H., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*.

Barman, U., Das, A., Wagner, J., and Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Bird, S. (2020). Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519.

Bird, S. and Loper, E. (2004). NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Calude, A., Stevenson, L., Whaanga, H., and Keegan, T. T. (2020). The use of Māori words in National Science Challenge online discourse. *Journal of the Royal Society of New Zealand*, 50(4):491–508.

Cho, K., van Merrienboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*.

Dijkstra, T. (2007). Task and context effects in bilingual lexical processing. In *Cognitive aspects of bilingualism*, pages 213–235. Springer.

Dunn, J. and Nijhof, W. (2022). Language identification for Austronesian languages.

Finn, A., Jones, P.-L., Mahelona, K., Duncan, S., and Leoni, G. (2022). Developing a part-of-speech tagger for te reo Māori. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 93–98.

Hämäläinen, M. (2021). Endangered languages are not low-resourced! In Hämäläinen, M., Partanen, N., and Alnajjar, K., editors, *Multilingual Facilitation*, pages 1–11. Rootroo Ltd.

Harlow, R. (1993). Lexical expansion in Maori. *The Journal of the Polynesian Society*, 102(1):99–107.

Higgins, R. and Keane, B. (2015). Te reo Māori — the Māori language — language decline, 1900 to 1970s', Te Ara - the encyclopedia of New Zealand.

Holmes, J. and Wilson, N. (2017). *An introduction to sociolinguistics*. Routledge.

James, J., Shields, I., Berriman, R., Keegan, P. J., and Watson, C. I. (2020). Developing resources for te reo Māori text to speech synthesis system. In *International Conference on Text, Speech, and Dialogue*, pages 294–302. Springer.

James, J., Shields, I., Yogarajan, V., Keegan, P. J., Watson, C., Jones, P.-L., and Mahelona, K. (2022a). The development of a labelled te reo Māori-English bilingual database for language technology.

James, J., Yogarajan, V., Shields, I., Watson, C., Keegan, P., Jones, P.-L.,

and Mahelona, K. (2022b). Language models for code-switch detection of te reo Māori and English in a low-resource setting. In *2022 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Jose, N., Chakravarthi, B. R., Suryawanshi, S., Sherly, E., and McCrae, J. P. (2020). A survey of current datasets for code-switching research. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 136–141. IEEE.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models.

Joulin, A., Grave, E., and Mikolov, P. B. T. (2017). Bag of tricks for efficient text classification. *EACL 2017*, pages 427–431.

Keegan, T. T. (2017). Machine translation for te reo Māori. In Whaanga, H., Keegan, T. T., and Apperley, M., editors, *He Whare Hangarau Māori Language, Culture & Technology*, pages 23–28. Te Pua Wānanga ki te Ao/Faculty of Māori and Indigenous Studies.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84.

Marras Tate, J. and Rapatahana, V. (2022). Māori ways of speaking: Code-switching in parliamentary discourse, Māori and river identity, and the power of Kaitiakitanga for conservation. *Journal of International and Intercultural Communication*, pages 1–22.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49.

Moorfield, J. C. (2021). Te aka Māori dictionary. `https://maoridictionary.co.nz/`.

Rapatahana, V. (2017). English language as thief. In *Language and Globalization*, pages 64–76. Routledge.

Scannell, K. (2011). Indigenous Tweets. `http://indigenoustweets.com/`.

Smith, L. T. (2021). *Decolonizing methodologies: Research and indigenous peoples*. Bloomsbury Publishing, third edition.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Te Hiku Media (2019). Identify Māori text. `https://github.com/TeHikuMedia/nga-kupu`.

Te Hiku Media (2022). reo-toolkit. `https://github.com/TeHikuMedia/reo-toolkit`.

Trye, D., Calude, A. S., Bravo-Marquez, F., and Keegan, T. T. (2020). Hybrid hashtags: #YouKnowYoureAKiwiWhen your tweet contains Māori and English. *Frontiers in artificial intelligence*, 3:15.

Trye, D., Calude, A. S., Bravo-Marquez, F., and Keegan, T. T. A. G. (2019). Māori loanwords: A corpus of New Zealand English tweets. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 136–142.

Trye, D., Keegan, T. T., Mato, P., and Apperley, M. (2022). Harnessing Indigenous Tweets: The Reo Māori Twitter corpus. *Language resources and evaluation*, pages 1–40.

Trye et al. (2022a). Hybrid architecture for labelling bilingual Māori-English tweets. `https://github.com/bilingual-MET/hybrid`.

Trye et al. (2022b). Interactive error analysis. `https://bilingual-met.github.io/hybrid/sample`.

Trye et al. (2022c). MET corpus explorer. `https://bilingual-met.github.io/hybrid/`.

White, T. H. (2016). A difference of perspective? Māori members of parliament and te ao Māori in parliament. *Political Science*, 68(2):175–191.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017). Understanding deep learning requires re-thinking generalization. In *Proc. International Conference on Learning Representations 2017*, pages 1–15.

Chapter 8

Analysing A/O Possession in Māori-Language Tweets

This chapter presents a case study of grammatical possession in Māori-language tweets. For practical reasons, our analysis is limited to examples where the possessive preposition *a* or *o* occurs between two nouns, which is a frequent construction in Māori. We employ the visualisation techniques presented in Part II—namely, the Heatmap Matrix Explorer (Chapter 4) and MultiCat (Chapter 5)—to analyse a curated subset of the RMT Corpus (Chapter 6). These techniques are used to explore semantic characteristics of possessive phrases, as well as the sociolinguistic profile of the people who wrote them. A key objective of this chapter is to demonstrate the kinds of insights that can be gained by applying the proposed visualisation techniques to a specific domain (corpus linguistics), thereby addressing the second research question posed in Chapter 1.

Publication Details

The following paper has been reproduced with minor changes to the formatting, as discussed in Section 1.4:

Trye, D., Calude, A. S., Harlow, R., & Keegan, T. T. (2024). Analysing A/O possession in Māori-language tweets. *Languages*, *9*(8), 271. https://doi.org/10.3390/languages9080271

Abstract

This article contributes the first corpus-based study of possession in Māori, the indigenous language of Aotearoa New Zealand. Like most Polynesian languages, Māori has a dual possessive system involving a choice between the so-called A and O categories. While Māori grammars describe these categories in terms of the inherent semantic relationship between the possessum and possessor, there have been no large-scale corpus analyses demonstrating their use in natural contexts. Social media provide invaluable opportunities for such linguistic studies, capturing contemporary language use while alleviating the burden of gathering data through traditional means. We operationalise semantic distinctions to investigate possession in Māori-language tweets, focusing on the [possessum *a/o* possessor] construction (e.g., *te tīmatanga o te wiki* 'the beginning of the week'). In our corpus comprising 2500 tweets produced by more than 200 individuals, we find that users leverage a wide array of noun types encompassing many different semantic relationships. We observe not only the expected predominance of the O category, but also a tendency for examples described by Māori grammars as A-marked to instead be O-marked (59%). Although the A category persists in the corpus, our findings suggest that language change could be underway. Our primary dataset can be explored interactively online.

8.1 Introduction

Like most of its Polynesian cousins, Māori, the indigenous language of Aotearoa New Zealand, has a complex possessive system involving a choice between the so-called A and O categories. This system encompasses a range of different forms (Harlow, 2000; Bauer et al., 1997, pp. 390–407), with A/O alternation having been described as one of the "thorniest" aspects of Māori grammar (Harlow, 2007, p. 170). Indeed, the A/O distinction presents a significant challenge for linguists, teachers, and learners alike (Fusi, 1985; Bauer et al., 1997; Thornton, 1998; Harlow, 2000).

This paper presents the first empirical analysis of this phenomenon in Māori by drawing on an existing corpus of Māori-language tweets (Trye et al., 2022), which we believe is representative of the language used by contemporary speakers of Māori. The paper has two main aims. Firstly, we intend to show how possessive markers are used by contemporary Māori speakers, documenting the extent to which this usage conforms with traditional descriptions of A/O alternation in Māori grammars. Secondly, because many corpus linguistic studies

involve categorical variables (Levshina, 2015; Stefanowitsch, 2020), we hope to demonstrate the general applicability of two novel visualisation techniques called *MultiCat* (Trye et al., 2024) and the *Heatmap Matrix Explorer* (Trye et al., 2023). Before outlining the details of this study, a general introduction to the topic of possessives is in order, together with some background information about the Māori language.

So important is the notion of possession that all languages have some way of expressing it (Aikhenvald, 2013, p. 1). Unsurprisingly, a wealth of accounts of linguistic possession has emerged, documenting the various systems found in the languages of the world. However, since it is neither possible nor desirable to summarise this body of work in only a few paragraphs, we limit our discussion to the most relevant points. The first observation is that, depending on the language, possessives can be expressed by a diverse array of constructions, including verbs, pronouns, case markers, and adpositions.

At its core, linguistic possession concerns relationships between entities. Here, we focus specifically on noun phrases encoding possession, also known as *adnominal possessives* (Haspelmath, 2017, p. 196). Possession involves a semantic relationship between a *possessor*, an entity that owns or possesses something, and a *possessum*[1] (pl. *possessa*), the entity owned or possessed. However, the notion of possession is fuzzy, occurring in figurative expressions that do not encompass a strict relationship of ownership, such as part–whole relationships (*fingers of the hand*, *pages of a book*), human relationships (*my friend's brother*, *John's wife*), features (*the cruelty of the war*, *the purpose of those canoes*), and cognitive processes (*my mother's thoughts*, *the boy's anger*), among others.

In considering possessive constructions, it is useful to separate the semantic relationship between the possessum and possessor on the one hand, and the grammatical form used by languages to express this relationship (which we loosely term 'possession') on the other. A frequent distinction is made between *alienable* and *inalienable* possession.[2] At the heart of this distinction lies the observation that certain possessive relationships involve situations in which the possessum and possessor are inextricably linked. For example, the notion of a biological *brother* implies the existence of at least one sibling, and body parts like *hand* and *mouth* are associated with the person to whom they belong. Haspelmath (2017, p. 197; 2021, p. 618) defines inalienable possessives as those involving body parts or kinship relationships. Some scholars have claimed

[1]Throughout this article, we refer to the possessed item as the *possessum*, rather than the *possessee*.

[2]See `https://grambank.clld.org/parameters/GB059#2/7.7/254.7`

there is a distinction between alienable and inalienable possession in Māori (e.g., Krupa, 1964, p. 434), with alienable possessa being marked by the A category and inalienable possessa by the O category; however, this has been superseded by more recent proposals based on the notions of dominance and control (Bauer et al., 1997, p. 390; Harlow, 2000, p. 363; 2015, p. 141).

An important dimension when analysing possessive marking systems, and one that is deemed highly relevant for Māori (see Section 8.2), involves finer distinctions between the various semantic classes of the entities involved. Semantic classes contain lexically specified sets of nouns with particular characteristics, which, in the case of Māori, and, indeed, most other Polynesian languages (Wilson, 1982, p. 3), attract different possessive markers. While grouping nouns into semantic classes entails knowledge of various language-internal idiosyncrasies, the vast amount of typological work on possession points to some common semantic classes relevant to possessive marking that recur across languages. For example, Chousou-Polydouri et al. (2023) identify the following classes: animals (sometimes split into wild or domesticated), body, kin (sometimes split into nuclear family, blood relations, or relations of marriage), inanimate natural entities, strict owner relations (owner, master), part–whole relationships, plants, place-related terms (native land, country), intimate property (furniture, tools, ornaments), names of people and places, mass nouns, and a mixed class with no semantic patterning. These categories provide a useful frame of reference for our analysis of Māori possessives (see Section 8.4.1).

Having introduced the phenomenon investigated in this article, we now provide some context about the language analysed. Our study focuses on possessive constructions in Māori, a Polynesian language spoken in the Pacific, in Aotearoa New Zealand. Māori is spoken by approximately 4% of the total population (Statistics NZ, 2018), including roughly one in six Māori adults (Te Kupenga, 2020). Most speakers acquire Māori either as compound or coordinate bilinguals, or as adult L2 speakers (King, 2018). Unsurprisingly, L1 and L2 speakers' language use differs considerably (Kelly, 2014; Christensen, 2003; Lane, 2024). Moreover, far from being a single, uniform language, Māori encompasses several distinct dialects (Harlow, 2007), including some variations concerning the use of the A/O categories (Biggs, 1955, p. 341). As is often the case with indigenous languages, the story of te reo Māori (the Māori language) is characterised by the Māori people's enduring fight for the survival and revitalisation of their language in the face of intense and ongoing colonisation (Greensill et al., 2017; Higgins et al., 2014; Whaanga and Greensill,

2014). One consequence of this language acquisition context is that English, as the dominant lingua franca in Aotearoa New Zealand, has had a significant impact on contemporary usage of Māori (Tawhara, 2015; Harlow, 2007; Harlow et al., 2011). Today, Māori is considered to be both endangered and low-resourced, though several corpora and digital tools have been developed to aid its revitalisation (for an overview, see Trye et al., 2022, pp. 1234–1236).

8.1.1 Scope

Due to limitations of time and space, our analysis focuses on a constrained but salient part of the possessive system, namely A/O alternation as found in constructions of the type [possessum *a/o* possessor].[3] In this construction, the possessum and possessor are both noun phrases, while the possessive marker functions as a preposition meaning 'of'. The *a/o* marker is used to "introduce possessive comments following nouns preceded by any determiner except *he*" (Harlow, 2015, p. 140).

Our rationale for focusing exclusively on the [possessum *a/o* possessor] construction is that it involves a clear dichotomy between the A/O categories, since there is no available neutral form (see Harlow, 2000, p. 365), and we thought it would be relatively straightforward to identify (compared to, say, the pairs *nā/nō* and *mā/mō*, which have numerous functions; see Harlow, 2015, pp. 72–73).

It is important to emphasise that the [possessum *a/o* possessor] construction relates to just one of five sets of Māori possessive particles, all of which involve A/O alternation (Harlow, 2000, p. 362). Nevertheless, we hope our target construction will serve as a microcosm for the wider possessive system, providing general insights into how Māori-language tweeters use the A and O categories. To the best of our knowledge, this research constitutes the first corpus-based study of possession in Māori, and the first study of any aspect of Māori grammar on social media. Besides the work of Kelly (2015) and Nicholas (2010), we are not aware of any in-depth quantitative research on naturally occurring Māori grammar. Our work aims to help bridge this gap by analysing a sizable Twitter[4] corpus of real-world data.

[3]By way of convention, we will use small letters in italics (*a/o*) to refer to the alternation between (single-vowel) *a* or *o* forms in our target construction, and non-italicised, capital letters (A/O) to represent the categories within the entire possessive system.

[4]Since our data were collected prior to Twitter's rebranding as X, we refer to the platform by its original name.

8.2 A/O Alternation in Māori

Like most Polynesian languages, possession in Māori entails a choice between the A and O categories (Harlow, 2000, p. 357). This distinction is the bugbear of many Māori-language learners, with some taking to social media to humorously share their frustrations ("Figuring out A and O is like doing long division in my head") and console themselves ("at least there's a 50/50 chance of getting it right"; comments on an Instagram post, 14 March 2024). Bauer et al. (1993, p. 209) note that "the 'correct' use of the A/O distinction is regarded as a shibboleth by many Māori speakers", which still frequently generates discussion as to why one form is preferable over the other.

According to Māori grammars, A/O alternation is contingent on the relationship between possessor and possessum (Harlow, 2015, p. 141; Bauer et al., 1997, p. 390). As a rule of thumb, the A class is best viewed as a 'special' (marked) relationship between the two (Clark, 1976, pp. 42–44), used in situations where the possessor is dominant (Bauer et al., 1997, p. 390) or the possessum comes under its protection or authority (Head, 1989, p. 102). This latter framing places responsibility on the possessor to be a good *kaitiaki* 'guardian'. The O class, on the other hand, is the 'default' (unmarked) category that applies in all other cases (Bauer et al., 1997, p. 391; Harlow, 2007, p. 168).

A wealth of proposals concerning the A/O categories has emerged over the past few decades, reflecting the complexity of this topic. As Fusi (1985, p. 119) puts it, "grammars try in different ways to give a satisfactory explanation ... and in the end all of them seem more or less (and each one in its own way) to have achieved an approximate comprehension of the mechanism involved, but without managing to give an exhaustive explanation of it." This variability applies to both the generalisations used, as well as explanations of individual exceptions. For example, the following labels have been posited by various scholars to succinctly capture the contrast between the A and O classes, respectively:

- alienability vs. inalienability (Krupa, 1964, p. 434; 2003, p. 122)
- active vs passive (Foster, 1987)
- dominance vs. subordination (Biggs, 1996, p. 42), later revised to dominance vs. non-dominance (Bauer et al., 1997, p. 391; Biggs, 2000, as cited in Harlow, 2007, p. 168)
- inheritance vs. active production (Ryan, 1974, p. 5)
- control vs. non-control (Moorfield, 1988, p. 140; Ryan, 1974, p. 75; Capell, 1949)

- higher vs. lower *tapu* 'potentiality for power' and *mana* 'prestige' (Thornton, 1998)

While, on the surface, these labels appear to be quite different, they all boil down to the notion of *agency* on the part of the possessor (even Krupa perceives the A category as dominant in his explanation of alienability vs. inalienability). Thornton (1998, p. 390) acknowledges that her categorisation with respect to *tapu* and *mana* is not robust for certain constructions, like subjects of nominalisations, which are "best described in grammatical terms" (see 3 below).

As mentioned above, our focus in this paper is on constructions of the form [possessum *a/o* possessor]. We provide three examples by way of introduction.[5] Examples (1-2) illustrate a distinction in the possessive system with respect to kinship ties, and constitute typical examples that a learner of Māori might be exposed to in beginner classes. In (1), an *a* marker is used because children are generationally below their parents; conversely, an *o* marker is used in (2) because parents are generationally above their children. Example (3) features the subject (*Rāwiri* 'David') of the nominalisation of an active transitive verb (*tuhinga* 'writing'), for which an *a* marker is used.

(1) *ngā* *tamariki* ***a*** *te* *matua*
the.PL children POSS the.SG parent
'The children of the parent / the parent's children'

(2) *te* *matua* ***o*** *ngā* *tamariki*
the.SG parent POSS the.PL children
'The parent of the children / the children's parent'

(3) *te* *tuhinga* ***a*** *Rāwiri* *i* *tana* *reta*
the.SG writing POSS David OBJ his letter
'David's writing of his letter'

Grammars and other pedagogical texts (e.g., Harlow, 2007, pp. 166–167; Harlow, 2015, pp. 140-146; Head, 1989, pp. 101-116) typically explain A/O alternation by grouping items into semantic classes, such as those detailed in Table 8.1. While full treatment is normally given to both the A and O categories, since O is unmarked, theoretically, "one need only specify when the *a*-forms should be used" (Harlow, 2007, p. 168).

[5]Throughout the paper, we write possessive markers in bold and use the following glosses: OBJ 'indirect object', PL 'plural', POSS 'possessive marker', and SG 'singular'. Macrons (e.g., *ā*) denote long vowel sounds in Māori.

Table 8.1: Summary of categories given in (Harlow, 2007, pp. 166–167).

A Class (Marked)	**O Class (Unmarked)**
Small portable possessions	Large objects, and animals used for transport
Kin of lower generations (see Example 1; apart from *uri*, 'descendant'), and spouses	Kin of same or higher generations (see Example 2)
Subjects of nominalisations of active transitive verbs (see Example 3), including derived nominals	Subjects of nominalisations of other verbs
Consumables, apart from water and medicine	*Wai* 'water' and *rongoā* 'medicine'
Animals not used for transport	Parts of whole, including body parts and clothing

Despite this last observation, a popular model used for teaching purposes is *Ngā Kawekawe o te Wheke* 'The Tentacles of the Octopus',[6] which inverts this approach by giving primacy to the O class. Developed by Pānia Papa and Leon Heketū Blake, this model specifies eight semantic categories, one for each tentacle, which take O forms rather than A forms. Generally, anything that falls outside these categories is likely to be A-possessed. To enhance memorability, the categories are all 'W' or 'Wh' words, which are salient sounds in Māori:

1. Whakarākei/Adornments
2. Whanaunga/Relations (of same generation or older)
3. Waka/Modes of transport
4. Wāhanga/Parts of someone/thing
5. Whakaruruhau/Shelter[7]
6. Whakaora/Wellbeing
7. Wāhi/Places
8. Whakaahua/Adjectives or qualities of someone/thing

While these kinds of lists (e.g., Table 8.1) and more general (e.g., control-based) explanations are helpful because they are practical for learners, as far as linguistic theory is concerned, they "miss what is going on at a more abstract level" (Harlow, 2007, p. 167). Unfortunately, exhaustive guidelines are not possible: there will always be cases involving noun phrases that are not explicitly covered by the specified rules (Kārena-Holmes, 2021). Some dictionaries list words as being either A-possessed or O-possessed (e.g., Williams

[6] `https://www.mumureo.com/product-page/ng-kawekawe-o-te-wheke-poster`

[7] This can include *people* who act as *kaitiaki* (e.g., doctor, teacher), as well as inanimate forms of shelter.

and Williams, 1971); however, such classifications are not definitive, since many noun phrases can occur in different contexts with a contrast in meaning (Fusi, 1985, p. 119). For example, a book or song *written* by someone takes A, while a book or song *about* someone takes O (Biggs, 1996, p. 42); clothes *designed* by someone take A, while clothes *owned* or *worn* by someone take O (Harlow, 2015, p. 141); a photo *taken* or *owned* by someone is A-possessed, while a photo *of* someone takes O (Harlow, 2000, p. 363). Unsurprisingly, learners often struggle with these fine-grained semantic distinctions, because they conceptualise the item in question as belonging to a specific category and, hence, having a fixed marker.

Adding to the confusion, lists of items for remembering the A/O categories inevitably contain exceptions or anomalies (Harlow, 2015, p. 141). Table 8.1 accounts for certain lexical exceptions (such as *uri*, *wai* and *rongoā*) and exclusions (e.g., animals not used for transport), but these are by no means exhaustive. For instance, contrary to Table 8.1, some intransitive verbs favour A (e.g., *noho* 'sit, stay') or can be used with both A and O forms, depending on what is being emphasised (Harlow, 2015, p. 188).

Control-based explanations of the A/O distinction also run into trouble with seemingly asymmetrical cases, such as the use of A for both one's husband (e.g., *te tāne a Mere* 'the husband of Mere') and one's wife (e.g., *te wahine a Tama* 'the wife of Tama'). While different explanations can be found (Harlow, 2007, p. 169), the point is that applying the notion of control (or similar) cannot straightforwardly account for all patterns observed. A further objection raised with such interpretations is that they are often presented from a Western perspective that does not always match the Māori worldview (Thornton, 1998; Fusi, 1985).

Another issue is speaker variation (Harlow, 2007, p. 168). Even proficient first-language speakers sometimes disagree about which marker should be used in a given context, and the same or different speakers may use contrasting markers without any perceptible difference in meaning (Harlow, 2015, p. 141; Bauer et al., 1997, pp. 393–394). Biggs (1955, p. 341) notes that A forms are more prevalent in the case of first-person possessors. Anecdotally, some speakers use A for drinking water, while others use O when referring to water used for both washing and drinking (Harlow, 2015, p. 143). Similar examples have been noted elsewhere in the literature (e.g., Fusi, 1985, p. 122).

A related problem is the ambiguity of classifications. Should a horse that is retired still be O-marked, as Table 8.1 might suggest, or should it be A-marked like other pets? Does it make a difference whether the horse in question

carries (or carried) *people* or *things*? There is no unanimous agreement here: both forms have been attested by native Māori speakers (Bauer et al., 1997, p. 393). Similarly, should a *hamarara* 'umbrella' be considered a small portable object (A) or a personal adornment (O), and does this change depending on whether it is in use? According to *Te Wheke* (the Octopus model), an umbrella that is being used provides a form of shelter (*Whakaruruhau*), which takes O, while an umbrella not in use is merely a *rākau* 'stick', which does not fit into any of the eight categories and, therefore, takes A. In these sorts of cases, a second-language speaker's use of A or O will again vary according to their understanding of the categories they are taught.

Furthermore, incoming words and changes in the environment affect the marking patterns used, sometimes also resulting in discrepancies between speakers. Bauer et al. (1997, p. 391) attribute many of the challenges with A/O classification to items introduced after European settlement, which disrupted the existing categories by partially fulfilling properties of each, and, therefore, not clearly fitting into either. For example, *waka*, being O-marked, was once the only form of transport, and generally belonged to the tribe rather than the individual. Bauer et al. (ibid) speculate that, when horses and cars came along, it was not clear whether they should be considered as items in an individual's control (and, hence, A-marked) or items that shared the same function as canoes (and, hence, O-marked), though most items of this nature were ultimately subsumed by the unmarked O category. As Krupa (2003, p. 133) notes, "The values valid within a community may undergo modification in time and, besides, conflicts in the classification as well as ambiguity no doubt stimulate changes and further evolution."

It is no wonder, then, that A/O alternation constitutes "one of the most complex aspects of Māori grammar" (Harlow, 2015, p. 140) and appears to have generated "more pages of text ... than any other single grammatical topic in Māori" (Bauer et al., 1997, p. 390).

We conclude this section with the personal perspective of one of the authors of the paper, who is an L2 Māori speaker, reflecting on his own experience acquiring possessive markers:

"As a second-language speaker of te reo Māori, I was introduced to the A and O categories early as another example of how *reo* (language) and *tikanga* (customs) are intertwined. To have confidence knowing which is the correct marker to use, one has to have a good understanding of the relationships being discussed and where the *mana* (authority) of the relationship resides. In general, the more authoritative actor has the O category, with the more

submissive character being referred to in the A category. But there are lots of nuances, with objects taking the O category in some instances, but then the same object taking the A category in other instances, depending on the different relationships that are in play. I was fortunate that when I learnt te reo Māori I was in the presence of some gifted exponents of Māori language. Once my ear became tuned, I found that I was able to defer to what sounded the best, rather than turning to specific rules about which category should be used. For me, this makes learning and speaking te reo Māori natural and more enjoyable as I am not relying on following rules, but rather I am being guided by a *taonga* (treasure) that has been handed down to me. Unfortunately, this does not mean I am correct all the time, but it works well enough for me to run with it!"

8.2.1 Research Questions

Given the complexity surrounding this topic, our work aims to investigate how contemporary Māori speakers handle A/O alternation on social media. Thus, our research questions are as follows:

- RQ1: What semantic categories and relationships are most frequently used by Māori-language tweeters in the [possessum *a/o* possessor] construction?
- RQ2: To what extent do Māori-language tweeters adhere to the rules described in Māori grammars when using the [possessum *a/o* possessor] construction? By analysing semantic categories and relationships, can patterns of adherence to and/or deviation from these rules be identified?
- RQ3: What are the sociolinguistic profiles of the tweeters in the corpus? If notable patterns arise from RQ2, can these be linked to characteristics such as gender, number of followers, and overall proportion of Māori-language tweets?

While these questions refer specifically to Māori-language tweeters, we believe the data are likely to be representative of everyday speakers of Māori in the current bilingual context, perhaps with a skew towards younger L2 speakers. Therefore, we aim to contribute not only to the understanding of Māori-language use on social media, but also contemporary speakers' use of Māori more generally.

We make the following hypotheses about each of our research questions. Based on research cited in Section 8.1 (e.g., Haspelmath, 2017), we hypothesise that constructions involving kinship terms and body parts will be the most

frequent in our data (RQ1). Secondly, with respect to RQ2, we anticipate that Māori-language tweeters, and by extension, contemporary speakers of Māori, will display a strong preference towards the O category, even in circumstances where this contravenes the usage described in Māori grammars. In terms of RQ3, we predict that women will deviate from this described usage more often than men, because they are known to be drivers of language change.

8.3 Data and Methods

The data for this study were sourced from the second release of the *Reo Māori Twitter (RMT) Corpus* (Trye et al., 2022). This dataset comprises 94,163 tweets posted by 2302 users over a 14-year period (2007-2022). For practical reasons, we focus on only a small subset of these tweets and users within our study (2484 tweets written by 237 users). All users in the RMT Corpus were originally identified through the *Indigenous Tweets* website (Scannell, 2022), and their tweets were obtained using the official Twitter API. To protect users' privacy, any links and usernames within each tweet were replaced with the placeholders <*link*> and <*user*>, respectively. Since the corpus predates Twitter's rebranding to X in July 2023, we refer to individual posts as 'tweets' throughout the paper.

Figure 8.1 summarises the process that we followed to prepare the data for analysis. Broadly, our approach consisted of two phases: an iterative selection process, aimed at curating a set of tweets that was relatively balanced across users, followed by an annotation procedure for determining whether the possessive marker used conformed with descriptions in Māori grammars. The steps within each phase comprised a mixture of computational and manual methods. Ultimately, we created two datasets for our study: a *Mixed Dataset*, which aimed to capture the 'real' distribution of A/O usage for our target construction, and an *A-Only Dataset*, whose purpose was to address the under-representation of *a* markers within the *Mixed Dataset*.

The first step in preparing the data was to systematically remove a large number of tweets that did not fulfil our requirements. Given the practical constraints associated with manual identification and annotation of tweets, we decided to focus on a small but meaningful subset of the RMT Corpus. Tweets were initially filtered out according to three criteria. First, we excluded tweets whose timestamp was unknown, as these would not be suitable for diachronic analysis. Furthermore, this helped to reduce the proportion of incomplete metadata, since many tweets that were missing timestamps were also missing

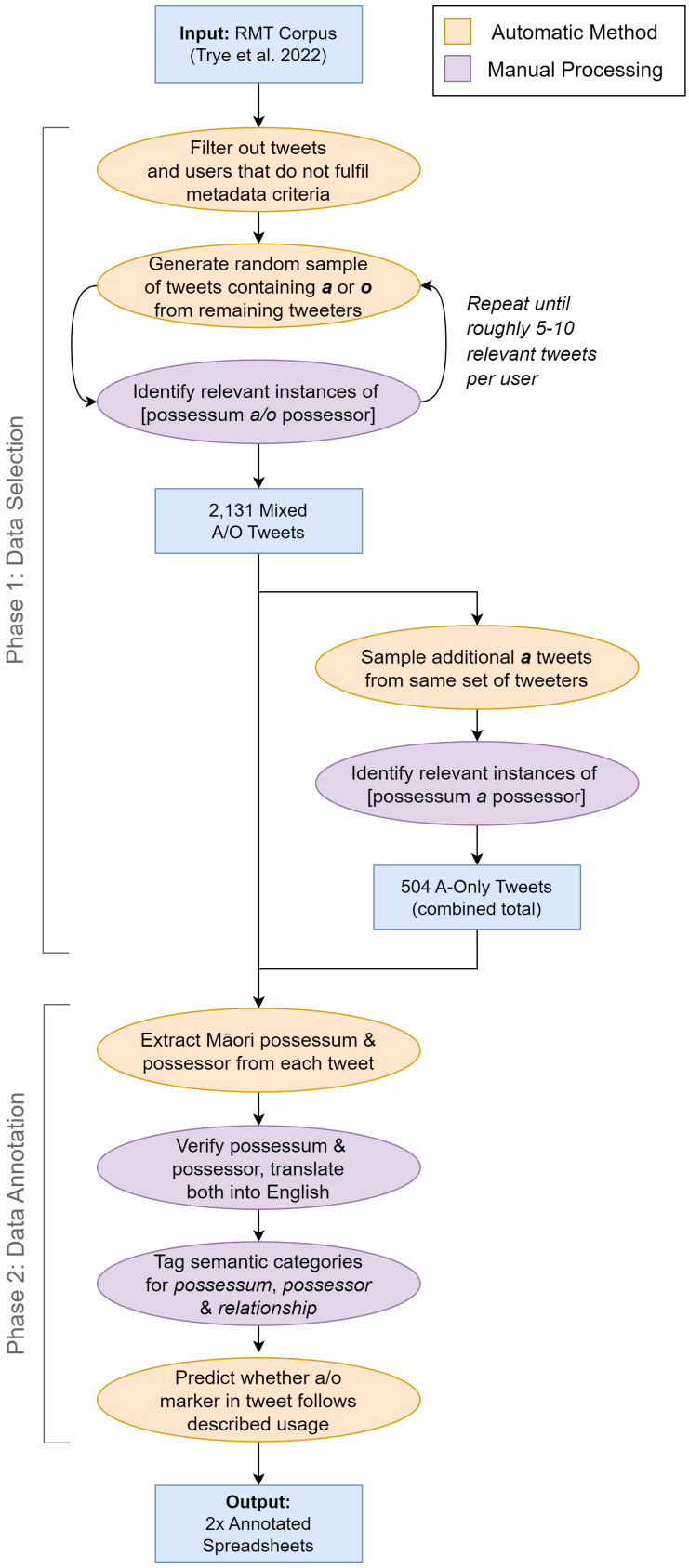


Figure 8.1: A visual overview of our data curation process. Some steps were performed computationally (orange), while others required manual processing (purple).

other values. Secondly, because we wanted each account to represent a discrete user, we removed tweets belonging to institutional and group accounts. Thirdly, we discarded tweets from users whose gender was unknown, as we wanted this to be the principal variable in our sociolinguistic analysis.[8] Overall, this step reduced the number of tweets in our sampling pool to 70,783 (a 25% decrease) and the number of users to 1582 (a 31% decrease).

From the remaining data, we generated a random sample comprising exactly ten tweets per user. We filtered out users whose combined number of tweets containing *a* and *o* (without macrons) fell short of this number. At this point, we also automatically removed specific instances of the *personal article* (Harlow, 2015, p. 67) and Cook Islands Māori, which is a separate language, by leveraging patterns in the data. If the same tweet contained multiple instances of *a* and/or *o*, we considered only the first instance. We limited the data to ten tweets per user because we did not want to over-represent prolific tweeters, opting instead for a more balanced and diverse user sample. This step reduced our user base by a further 85%, leaving only 241 users, and yielding an initial sample of 2410 tweets.

The next step involved manually scanning the tweets to determine whether they constituted relevant instances of the [possessum *a/o* possessor] construction. In particular, we removed the following irrelevant tweets:

- Non-possessive uses of *a* and *o*, including occurrences of the English indefinite article, incorrect word division (e.g., *a*'s that had become detached from *kia*), and remaining instances of the personal article;
- Tweets in which (short) *a* or *o* were used instead of (long) *ā* or *ō*, respectively;
- Formulaic phrases[9] in which users did not explicitly choose a possessive marker (e.g., *Te Wiki o te Reo Māori* 'Māori Language Week');
- Tweets where the use of *a* or *o* was unclear.

Manual inspection of the 2410 tweets in our first sample revealed that roughly a third of the data (780 tweets) were irrelevant. This included the overwhelming majority of tweets containing *a* (80%), as well as 20% of tweets containing *o*. After removing these tweets, most users in our dataset had fewer than ten tweets, and many had fewer than five. An iterative sampling strategy was employed to mitigate this imbalance across users.

[8]Information about the gender of each tweeter came from the RMT Corpus, and was primarily based on users' self-reported pronouns. We acknowledge that Twitter users may claim identities that they do not possess or wish to possess.

[9]There are still some formulaic names of entities in our data, but these represent only a small proportion of tweets and are not productively used.

Since many users had additional tweets containing *a* and *o* that were not part of our first sample, we randomly selected (from the leftover supply) the number of tweets needed to bring each user's total up to ten, or as close to ten as possible. The relevant tweets from the new sample were then added to those from the previous sample. This process was repeated three times, with each successive sample becoming smaller as we procured more relevant tweets from the necessary users and/or exhausted their leftover supply. To expedite the tagging process, we automatically removed any tweets containing recurrent formulaic phrases that were present in our first sample (e.g., *Te Tiriti o Waitangi* 'the Treaty of Waitangi' and *Te Wiki o te Reo Māori* 'Māori Language Week').

Unfortunately, even after completing our iterative sampling procedure, many users still had fewer than ten relevant instances of our target construction. Consequently, we settled on a final range of between four and ten tweets per user. We removed four users whose final tweet count fell below this threshold.

All of this resulted in our final *Mixed Dataset*, which comprised 151 *a*-marked tweets and 1980 *o*-marked tweets. The distribution of tweets per user is shown in Figure 8.2. Just over half of the 237 users (52%) had a complete set of ten tweets and over three-quarters (78%) had at least nine tweets. Overall, we manually checked 3297 tweets when creating the *Mixed Dataset* and retained 65% of these. The proportion of discarded tweets for each marker closely matched our first sample (80% for *a* and 22% for *o*), which is not surprising, given that the first sample was by far the largest.

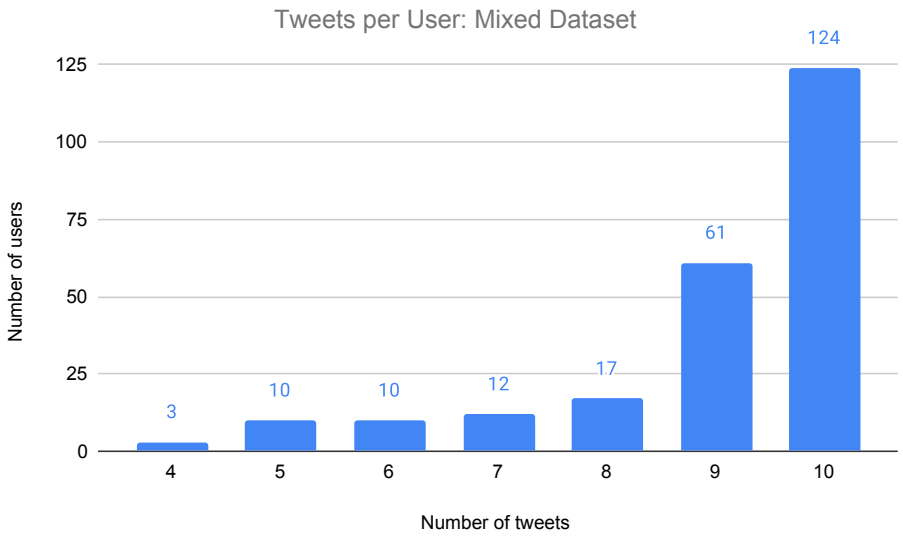


Figure 8.2: The number of tweets per user in the *Mixed Dataset*.

Due to the under-representation of the A category in the *Mixed Dataset*, we decided to gather additional instances of the *a* form so that we could more thoroughly examine its use on social media. To this end, we generated a further random sample of 1640 tweets from users in the *Mixed Dataset* who had leftover tweets containing *a*. We manually checked this sample and removed irrelevant instances in the same manner as that for the *Mixed Dataset*. However, when sampling the data, we extended the acceptable range of tweets per user to between 1 and 20, recognising that, because *a* is relatively infrequent as a possessive marker, we would need to draw more heavily on prolific tweeters in the corpus. Only 353 of the tweets in the sample (22%) were relevant. We combined these with all instances of *a* from the *Mixed Dataset*, resulting in 504 *a*-marked tweets in total. We refer to this as the *A-Only Dataset*. We chose not to merge our two datasets completely, because we did not want to interfere with the natural distribution of *a/o* markers in the *Mixed Dataset*.

The number of tweets per user in the *A-Only Dataset* is shown in Figure 8.3. The dataset was skewed in the opposite direction to the *Mixed Dataset*: due to the lack of data, there were more users with fewer tweets. Indeed, most users have fewer than five relevant *a* tweets. The dataset contains tweets from 58% of users (137 of 237) in the *Mixed Dataset*, which shows that most users did not disregard *a* altogether, even if they used it much less frequently than *o*. However, there is still a large proportion of users (42%) in the *Mixed Dataset* that did not use *a* at all.

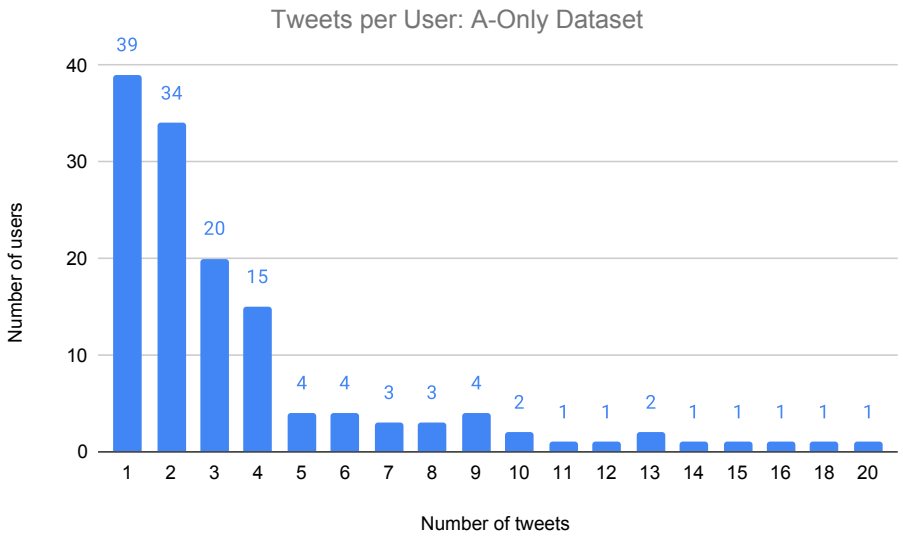


Figure 8.3: The number of tweets per user in the *A-Only Dataset*.

Having gathered the data for our analysis, we were now in a position to annotate them. We attempted to computationally extract both the possessum and possessor based on their positions. Following this, an L1 Māori-language speaker manually updated any items that had not been correctly identified. The same speaker then translated the possessum and possessor from Māori into English, using the definition they deemed most appropriate within the context of each phrase and consulting the online Māori dictionary (`https://maoridictionary.co.nz/`) where necessary. Given that many words are polysemous in Māori(Boyce, 2006), this was not a straightforward task. We deleted a handful of tweets where the possessum or possessor was neither familiar to the speaker nor listed in the Māori dictionary (e.g., *te rauiwi*, *te tikene*).

One of the authors then tagged each possessive phrase according to three semantic variables (*PSSM, PSSR* and *RELA*), as detailed in Section 8.4. Briefly, the *PSSM* and *PSSR* variables incorporate the same set of 22 semantic categories to classify the possessum and possessor, respectively, while the *RELA* variable comprises a distinct set of 13 categories for encoding the relationship between the two. The author responsible for assigning the semantic categories flagged any tweets they were unsure about, which were then checked by another author with greater proficiency in Māori. We discarded a small proportion of tweets which, despite initially having been marked as 'relevant', were re-evaluated as formulaic or ambiguous.

Since most of the semantic coding was done by the same person, we sampled 10% of the data for another analyst to code independently. The inter-rater reliability scores for this sample are given in Table 8.2, showing strong overall agreement for all three semantic variables.

Finally, we created a variable called *Type*, which drew on information from Māori grammars to predict whether each possessive phrase in our data used the described, and, therefore, expected marker. Further details are given in Section 8.4.4.

Table 8.2: Inter-rater reliability for all three manually coded semantic variables.

Variable	**Cohen's Kappa**	**Interpretation**
PSSM	0.89	Near perfect
PSSR	0.88	Near perfect
RELA	0.79	Substantial

8.4 Semantic Classification Scheme

Since our semantic classification scheme is crucial to our analysis, we provide further context about the development of these categories and the main challenges encountered. As mentioned in Section 8.3, we coded separate variables for the possessum (*PSSM*), possessor (*PSSR*), and the relationship (*RELA*) between the two. The literature on Māori suggests that our *RELA* variable is the principal factor determining the choice of possessive marker (Harlow, 2000, p. 363; Bauer et al., 1997, p. 290; Harlow, 2015, p. 141). However, we still wanted to code the other two variables, as we were interested in exploring the most common kinds of nouns used in possessive phrases and uncovering associations between the three variables (for example, do certain relationships predominantly occur with specific types of possessum or possessor?). We devised our own classification scheme for these variables, drawing on existing Māori grammars and guides, especially Head (1989), Biggs (1996) and Harlow (2015). Our process for fine-tuning the labels was highly collaborative and resulted from several coding sessions where we discussed actual examples from our data.

8.4.1 *PSSM* and *PSSR* Variables

The 22 categories that we devised for the *PSSM* and *PSSR* variables are shown on the right side of Figure 8.4, while full definitions and examples are given in Appendix F (Table S1). One of the difficulties we faced was deciding on an appropriate number of categories. Our goal was to strike a sensible balance by capturing the intricacies of A/O alternation in Māori, without making the categories so specific that they applied to only a handful of noun phrases. To give just one example, we felt there is an important distinction between intrinsic components of cultural identity, such as language and religion, and deliberate forms of creative expression by groups or individuals, such as music, poetry, and dance. This led us to create two separate categories for cultural artefacts, namely *cultural_foundation* and *cultural_artistry*.

As alluded to in Section 8.1, our semantic categories share obvious similarities with those developed by Chousou-Polydouri et al. (2023, Table 4, p. 1382) for the purpose of cross-linguistic comparison. Figure 8.4 provides a visual overview of similarities and differences between their framework (shown in orange) and our own (shown in blue), which was created independently. Categories that are vertically aligned within the black rectangle exhibit substantial overlap, though they are not necessarily identical. For example, the orange

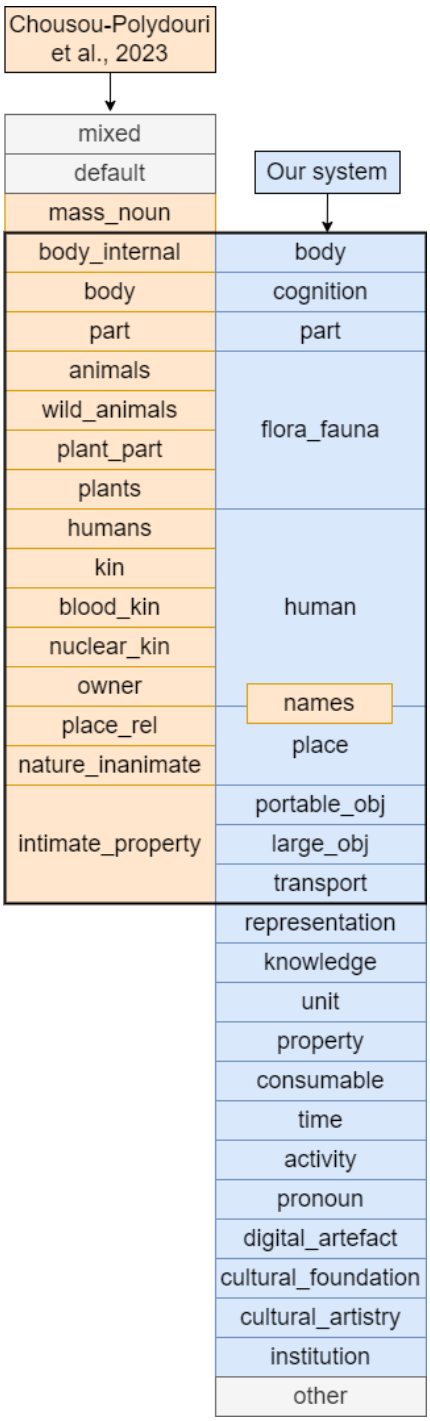


Figure 8.4: Overview of semantic categories used in existing research and in the present study. Grey boxes denote classes that are not semantically coherent.

body category, distinct from *body_internal*, covers mental faculties and feelings, and, therefore, aligns closely with our *cognition* category. As one might expect, our language-specific system comprises several categories that do not correspond with Chousou-Polydouri et al.'s more general approach, while the reverse is not true. In most cases, these extra categories were incorporated because they were mentioned in Māori grammars and/or occurred multiple times in our data.

8.4.2 *RELA* Variable

As noted above, the literature on Māori emphasises that the *relationship* between possessum and possessor is of paramount importance. For this reason, we coded a separate variable, called *RELA,* with 13 kinds of relationships. Detailed definitions and examples are given in Appendix F (Table S2), but we provide a summary here. We created four categories for interpersonal relationships (i.e., possessive phrases where both the possessum and possessor are human): *<kin, <non-kin, >=kin* and *>=non-kin.* As the names suggest, the appropriate category depended on whether the possessum and possessor had a familial connection, and on whether the former was senior to, or responsible for, the latter. We distinguished between an *ownership* category for personal belongings/assets, and a *creation* category for entities made or produced by the possessor. A *creation/ownership* category was reserved for cases where there was not enough context to determine which of these relationships applied. We used a *partitive* category for part-whole relationships, whereby the possessum required a possessor, though possibly implied. For example, one cannot be a *mema* 'member' without being a member of *something* (e.g., *te rōpū* 'the group'). In contrast, in a *descriptor* relationship, the possessum could occur independently of the possessor, but was specified, limited, or defined by it (e.g., 'the Queen (of England)'). A *feature* relationship denoted cases where the possessum was a quality of the possessor, such as an abstract noun, while *representation* applied when the possessum symbolically, nominally or visually represented the possessor. Finally, the *nom_agentive* and *nom_other* categories were used for nominalisations of different kinds of verbs. They can be seen as grammatical categories that override semantic rules, though the notion of agency still applies. The same verb form could occur in the *nom_agentive* or *nom_other* categories depending, for instance, on whether active or passive voice was used (Biggs, 1996, p. 42).

8.4.3 Semantic Annotation Challenges

As with any semantic classification system, annotating the data involved a certain degree of subjectivity. We encountered four main challenges when coding our semantic variables, which are discussed below. All examples in this section provide the original Māori text, together with the semantic classes assigned and an idiomatic English translation.[10]

The first challenge we faced concerned the overlapping nature of semantic categories. Despite our efforts to delineate clear semantic boundaries, we found that certain noun phrases straddled multiple categories. In order to treat these items as consistently as possible, we added stipulations to our category definitions that were not evident from the name of the category alone (see Tables S1 and S2 in Appendix F). For example, we consistently classified all types of buildings as *place*, even though they are also immovable, man-made structures (*large_obj*). This applied to the noun phrases *whare* 'house', *kāinga* 'home', *marae* 'meeting house', *whare tapere* 'theatre', *hōhipera* 'hospital' and *whare pukapuka* 'library', among others. Similarly, geographical features such as *maunga* 'mountain' and *moana* 'sea' were coded as *place* rather than as *flora_fauna.* Items such as *kupu* 'word', *rerenga* 'sentence', *kōrero* 'speech' and *kīanga* 'idiom' were all coded as a *unit*, not *representation*, since they constitute linguistic units of an utterance or piece of text.

Secondly, even after adjusting our category definitions, we found that the labels for many noun phrases and relationships were highly context-dependent. We needed to consider each possessive phrase in its entirety, but in many cases, it was also necessary to examine the wider tweet. Noun phrases that occurred in multiple tweets were sometimes assigned different labels due to subtle differences in meaning. For example, *pukapuka* 'book' was categorised as *cultural_artistry* when referencing its literary content or author, while a physical copy of a book was labelled *portable_obj.* (Had it occurred in our data, an electronic copy would have been classified as a *digital_artefact.*) We wanted our system to reflect these nuances in meaning, in case they proved relevant for determining the appropriate possessive marker.

Moreover, as explained in Section 8.2, Māori grammars emphasise that the same noun phrase, and even the same possessum-possessor pair, does not guarantee the same kind of *relationship* (i.e., *RELA* category). Returning to the example of *pukapuka*, a book is involved in a *creation* relationship if the

[10]The examples are written as they appear in our data, reflecting each speaker's choice of possessive marker. Unexpected markers are prefixed with an asterisk. Semantic categories are given in the order <*PSSM*>, <*RELA*> and <*PSSR*>, with <*RELA*> appearing directly beneath the possessive marker.

possessor is the person who wrote the book or the company that published it; example (4) references Witi Ihimaera's book *Sleeps Standing Moetū* about the Battle of ōrākau, which was translated into Māori by Hēmi Kelly. In contrast, an *ownership* relationship is applicable if the possessor owns a copy of the book but did not write or publish it, and a *descriptor* relationship applies if the possessor is the subject matter of the book. At first glance, example (5) might appear to be about a book owned or produced by a (possibly hypothetical) 'C Company'. However, the rest of the tweet reveals important context: this particular book is written by New Zealand historian Monty Soutar, who—a quick Google search revealed—is the author of *Nga Tama Toa: The Price of Citizenship: C Company 28 (Maori) Battalion 1939-1945*. Therefore, (5) almost certainly references this non-fiction book *about* the Māori Battalion's C Company, which served during World War II. Following Māori grammars, a *creation* relationship, such as that in (4) is A-marked (as would be an *ownership* relationship), while a *descriptor* relationship, such as that in (5), is O-marked. These examples illustrate how context is crucial in deciphering the precise nature of a possessive relationship.

(4)	*te*	*pukapuka*	***a***	*Hēmi Kelly*	*rāua*	*ko*	*Witi Ihimaera*
	the.SG	book	POSS	Hēmi Kelly	both	with	Witi Ihimaera
	<cultural_artistry>		<creation>	<human>			
	'the book of (written by) Hēmi Kelly and Witi Ihimaera'						

(5)	*te*	*pukapuka*	***o***	*Kamupene C*
	the.SG	book	POSS	C Company
	<cultural_artistry>		<descriptor>	<institution>
	'the book of (about) C Company'			

A third challenge involved navigating polysemous words, which are prolific in Māori (Boyce, 2006). It was important to determine the correct sense of each word in our possessive phrases in order to identify the appropriate relationship. For instance, *tikanga* can mean 'custom; practice; tradition' (as shown in examples 6-7) or 'meaning' (8-9). In example (6), the tweeter discusses customs *created* by Pākehā (European New Zealanders), with which they are unfamiliar. Example (7) refers to the *tikanga* 'customs' that operate within the *kāinga* 'home' (*descriptor*), although it could also be argued that *kāinga* in this context is a personification of the *people* who created the *tikanga*, in which case a *creation* relationship would again be appropriate. When *tikanga* is used in the sense of 'meaning', as in (8), this most obviously fits within our *property* category and the associated *feature* relationship: the meaning of something is a property/feature of that thing. Upon reading (9), it is not clear whether

the user is asking what the initials "NZTL" stand for, or if they are enquiring about NZTL's practice. However, by viewing the tweet in context,[11] we found a reply that indicated it was the former ("New Zealand Twitter League").

(6)	*ngā*	*tikanga*	***a***	*Te*	*Pākehā*
	the.PL	customs	POSS	the.SG	European
	<cultural_foundation>		<creation>	<human>	

'the customs of the European(s)'

(7)	*ngā*	*tikanga*	***o***	*te*	*kāinga*
	the.PL	traditions	POSS	the.SG	home
	<cultural_foundation>		<descriptor>	<place>	

'the traditions of the home'

(8)	*te*	*tikanga*	***o***	*te*	*kupu*
	the.SG	meaning	POSS	the.SG	word
	<property>		<feature>	<unit>	

'the meaning of the word'

(9)	*BTW*	*he*	*aha*	*te*	*tikanga*	****a***	*#NZTL?*
	BTW	what	is	the.SG	meaning	POSS	#NZTL
				<property>		<feature>	<institution>

'B(y) T(he) W(ay,) what is the meaning of #NZTL?'

A related difficulty arose from the fact that Māori can mark nominalisations by zero morphs (Harlow, 2007, p. 27). In a number of cases, there also occur nouns of the same form, meaning the object produced by the particular verb. For example, *kōrero* can mean 'to speak' but also (by zero-derivation) 'story, speech, discourse'; *waiata* can mean 'to sing' but also 'song' (Harlow, 2007, p. 102). Since such nominalisations are indistinguishable from noun forms, it is important to determine whether one is referring to the *action* itself (a nominalised verb) or the *product* of the action (a noun). Examples (10) and (11) capture this distinction: (10) refers to the children's act of speaking, without telling us anything about their conversation (a *nom_agentive* relationship), whereas (11) comments on the accuracy of what Haimona said, involving the relationship of *creation*. While example (12) may seem similar to (11), it differs in that the acts of speech that the tweeter says should be celebrated are those *about* the country, not those *created by* the country.

[11] Even though links within tweets were redacted from the RMT Corpus, we still had access to the URLs for the tweets themselves and could, therefore, view them on Twitter if they were (still) publicly available.

(10) *i* *oho* *au* *waenganui* *i*
TENSE wake I the middle of
te *kōrero* ***a*** *āku* *tamariki*
the.SG conversation POSS my children
<activity> <nom_agentive> <human>
'I woke up in the middle of my children's conversation
[lit. the speaking of my children]'

(11) *E tika ana* *te* *kōrero* ***a*** *Haimona* *nei!*
correct the.SG speech POSS Simon PARTICLE
<unit> <creation> <human>
'What Haimona said [lit. the speech of Haimona] is correct!'

(12) *me* *whakanui* *tatou* *ngaa* *koorero* ***o*** *eenei* *motu*
should celebrate we the.PL speech POSS this country
<unit> <descriptor> <place>
'We should celebrate the [good] things said about this country'

A fourth challenge was the lack of sufficient context. Due to the nature of the data analysed, the context of certain possessive constructions was either unfamiliar to us or missing altogether. This problem was exacerbated by the general noisiness of social media data, as well as Twitter's character constraints. Despite our best efforts, there was sometimes not enough detail in a tweet for us to ascertain its intended meaning. In such cases, we made an educated guess about the most likely interpretation of the tweet.

Fortunately, in cases where the meaning of a tweet was not clear, it was sometimes possible to find additional contextual clues. Some tweets were part of a longer thread or discussion involving multiple tweeters (e.g., example 9), and some contained useful photos or external links. However, not all tweets were still available on Twitter, so this extra context was not always available to us. As we have already seen from examples (4) and (5), in some cases, we could search online to 'fill in the gaps'. This also proved helpful in determining whether proper nouns that were unfamiliar to us referred to a *human*, *place, institution* or *activity.*

8.4.4 *Type* Variable

The final step in preparing the data was to determine whether the possessive marker used in each tweet matched the usage described in Māori grammars; this was encoded in a variable called *Type*. First, we created an intermediate variable called *Predicted* to store the expected marker (*a* or *o*) for each possessive phrase. The algorithm for computing this value is shown in Figure 8.5 and consists of three steps:

1. Address lexical exceptions, as identified in Harlow (2007) and Head (1989). Of these items, only *uri* 'descendant', *wai* 'water' and *wahine* 'wife' occurred in our data.
2. Assign markers based on the *RELA* variable, but manually check the *creation/ownership* category and ignore *ownership* relationships. This is the most crucial step, as it applies to the largest proportion of data.
3. For all *ownership* relationships, assign markers based on ten *PSSM* categories with fixed A/O forms, and manually check the rest.

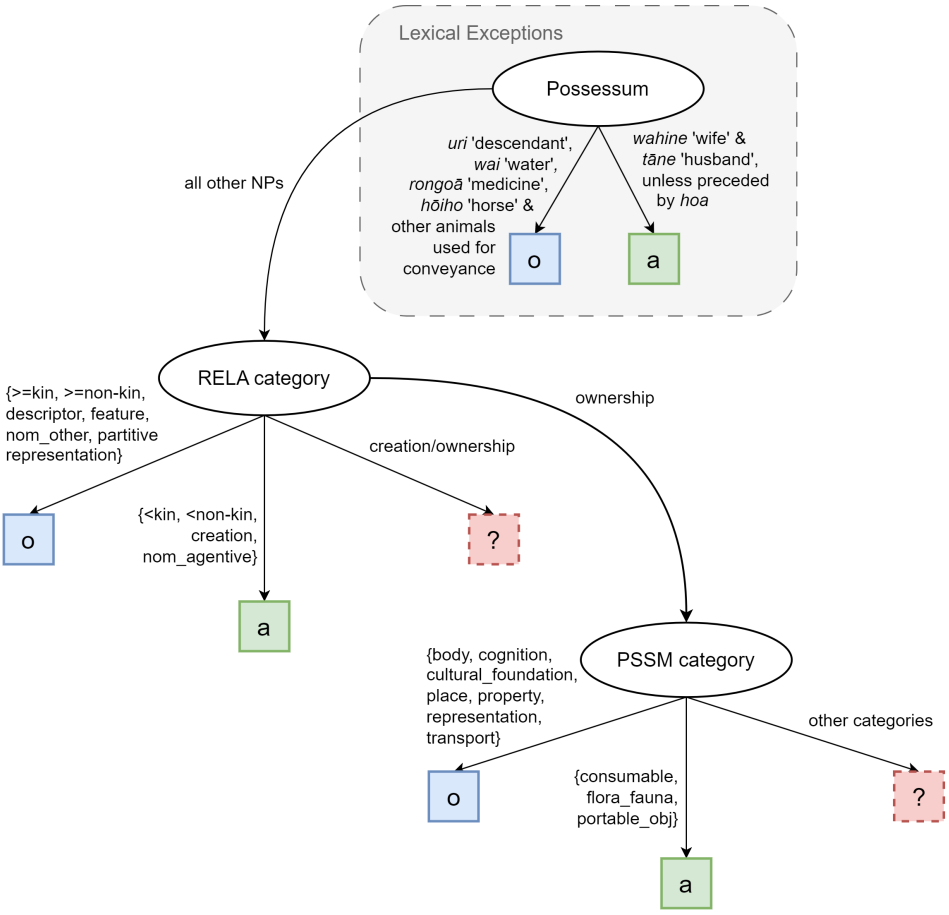


Figure 8.5: Our algorithm for determining the *Predicted* marker for each possessive phrase.

We then compared the assigned *Predicted* value for each possessive phrase against the marker actually used in the tweet, giving rise to our *Type* variable with the following four categories:

1. *a_expected*, if both the predicted and actual markers were *a*;
2. *a_unexpected*, if the predicted marker was *o* but the user chose *a*;
3. *o_expected*, if both the predicted and actual markers were *o*;
4. *o_unexpected*, if the predicted marker was *a* but the user chose *o*.

8.5 Results

We will now address each of our three research questions in turn. Our findings are presented using several visualisation methods, including two novel techniques for analysing categorical data: *MultiCat* (Trye et al., 2024) and the *Heatmap Matrix Explorer* (Trye et al., 2023). Traditional static plots are useful for showing overall frequency distributions, whereas these novel interactive techniques facilitate the exploration of more complex relationships in the data. Since the figures provided in the paper are necessarily static, we provide a link, `https://dgt12.github.io/possession/`, for readers to probe the *Mixed Dataset* themselves, as detailed in Section 8.5.1. This interactive capability is, in our view, one of the main benefits of using such techniques. All figures in this section relate to the *Mixed Dataset*, unless otherwise stated.

8.5.1 Semantic Variables by Frequency (RQ1)

We first investigate the frequency of the semantic categories and relationships in the data, examining individual categories before inspecting the patterns that arise when they are cross-tabulated. Our goal here is to shed light on the *kinds* of things that Māori-language tweeters typically discuss when using possessive phrases, regardless of whether they use the marker that conforms with the usage described in Māori grammars.

Figure 8.6 shows the frequency of the semantic categories in the *Mixed Dataset* for both possessa (*PSSM*, shown in blue) and possessors (*PSSR*, shown in red). All 22 categories are represented across both positions, although there are only 4 *part* possessors and 8 *transport* possessa. There are very few instances of *large_obj* or *portable_obj* in either position, even though they are broader in scope—in terms of candidate noun phrase 'types'—than several other categories (e.g., *transport* and *cultural_foundation*). Looking at the most frequent categories, there are many more *human*, *place, time*, and *institution* possessors than possessa. Conversely, the categories *activity*, *property, part*, and *unit* are more productive as possessa than possessors. The three most common possessum categories are *property, human*, and *activity*, which each have similar frequencies ($\tilde{1}7\%$) and together account for 50% of possessa. *Human, place*, and *time* are the three most common possessor categories (28%, 24%, and 12%, respectively), constituting 64% of possessors. The presence of the *human* category in both of these top-three rankings, and as the most frequent category overall, provides strong evidence that possessive phrases frequently relate to people.

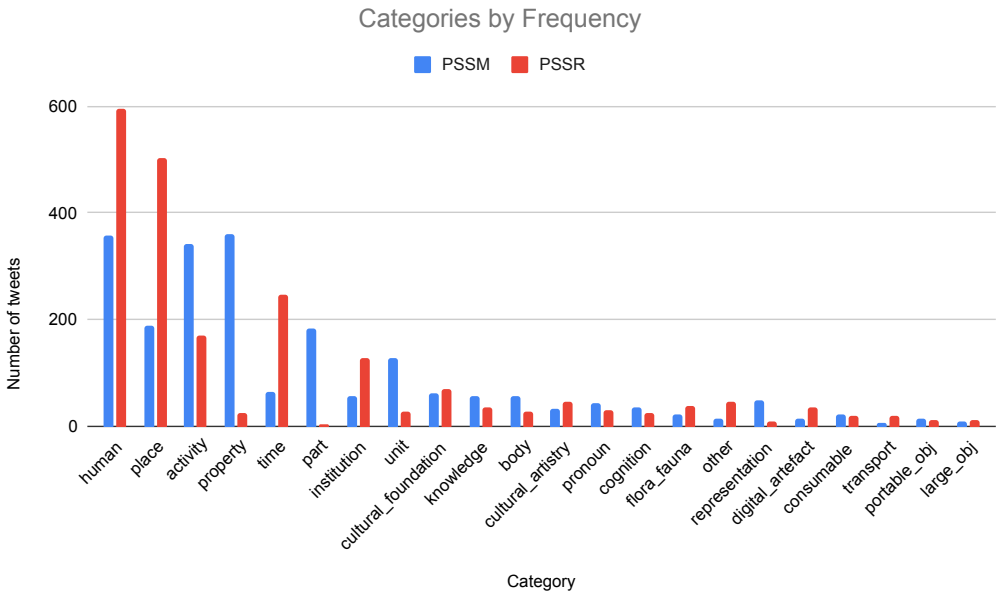


Figure 8.6: Grouped bar chart showing the frequency of each semantic category for both the possessum (blue) and possessor (red).

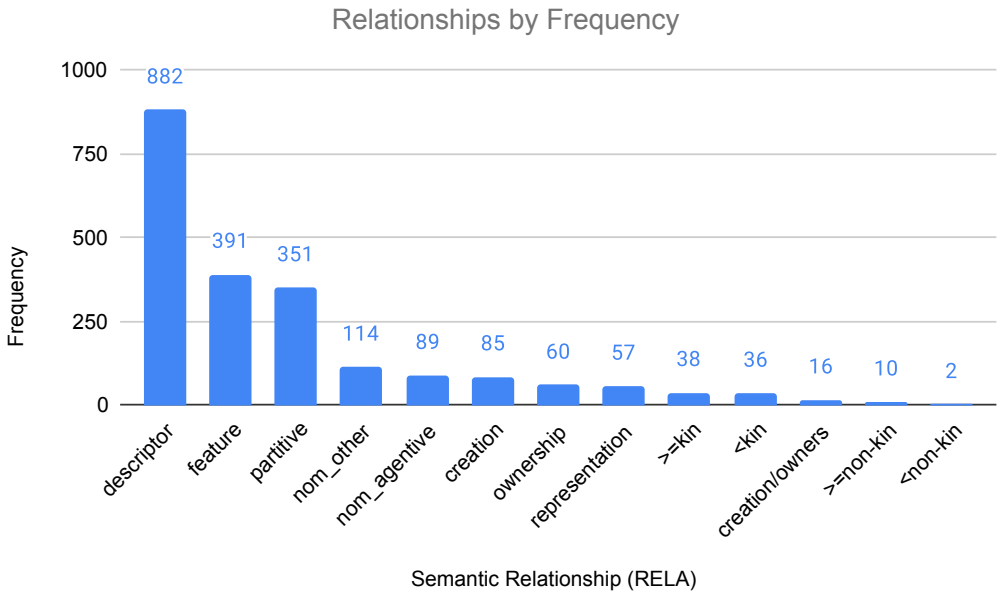


Figure 8.7: All 13 semantic relationships ordered by frequency.

Next, we analyse the frequency of the 13 semantic relationships in our data, as shown in Figure 8.7. *Descriptor* relationships are by far the most common, occurring more than twice as often as the next most frequent categories, *feature* and *partitive*. Both types of nominalisation (*nom_other* and *nom_agentive*) are ranked next—though *feature* and *partitive* relationships are more than three times as likely—followed by *creation, ownership*, and *representation* relationships. It is interesting that nominalisations are relatively frequent, as they tend to be presented peripherally or not at all in teaching resources and lists that assign the A/O categories to individual words. Kinship relationships (>=*kin*, <*kin*) are relatively infrequent, contrary to our hypothesis for RQ1. Finally, the least frequent categories are the two non-kin relationships (>=*non-kin*, <*non-kin*), which, together, occur only a dozen times. Therefore, it would appear that, while humans are very frequent as either a possessum or possessor (Figure 8.6), interpersonal relationships, involving a human in *both* slots, are much less common.

Looking at the noun phrases in our data, there are 1173 distinct possessa and 1227 distinct possessors in the *Mixed Dataset*. Roughly a quarter (276) of these possessa and a fifth (266) of these possessors occur more than once. Figure 8.8 shows noun phrases that appear at least 20 times in the *Mixed Dataset*, considering their use as both a possessum (blue) and possessor (red). When processing the data, we converted noun phrases to lowercase and removed macrons to consolidate similar items, but the same phrase may still have multiple forms due to spelling errors, dialect variations, and the use of

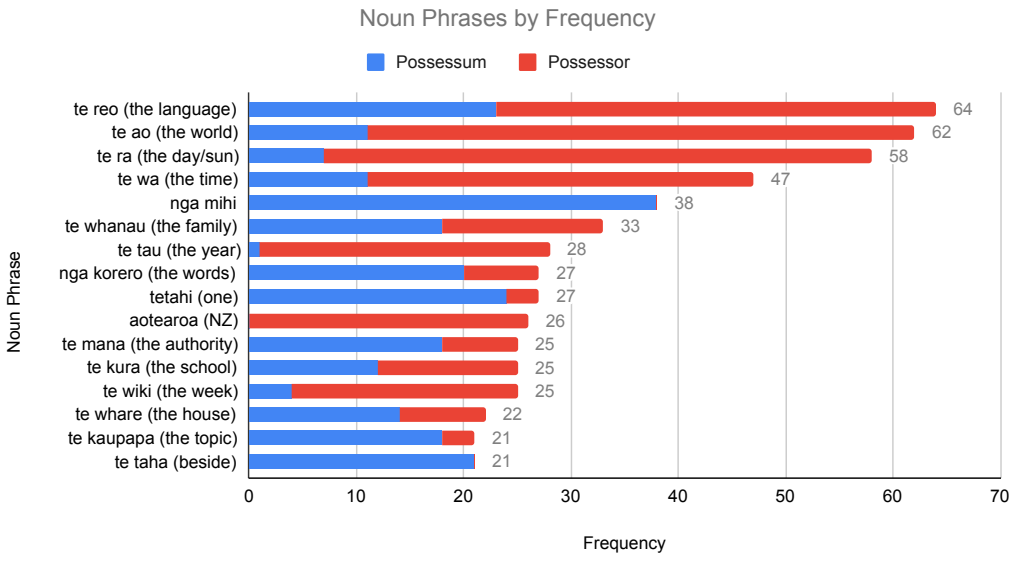


Figure 8.8: Noun phrases that appear at least 20 times in the *Mixed Dataset*.

double-vowel orthography. Note that some items are polysemous and, thus, not necessarily used with the same meaning. Most items in Figure 8.8 comprise a mixture of blue and red, showing that common noun phrases can and do occur in both positions,[12] although one particular use may be more common. For example, the four most prolific noun phrases, *te reo* 'the language', *te ao* 'the world', *te rā* 'the day/sun' and *te wā* 'the time', all occur predominantly as possessors. In contrast, the fifth most frequent noun phrase, *ngā mihi* 'acknowledgements', is used exclusively as a possessum, as is *te taha* 'the side', which appears at the bottom of the chart.

When considering the dataset as a whole, rather than just the most frequent noun phrases, the possessa that occur in the most distinct relationships are *ngā mahi* 'the deeds/work', *te tangi* 'the funeral', *te waka* 'the canoe', and *te whare* 'the house', which each occur in four distinct relationships. The possessors that occur in the most relationships are *te whānau* 'the family' (9 relationships), *te reo* 'the language (7 relationships), *te atua* 'the god' (6 relationships), and *te tangata* 'the people' (6 relationships). These statistics suggest that possessors tend to be used in more diverse ways than possessa.

Due to the nature of our classification system, there are some clear associations between the possessum (*PSSM*) and relationship (*RELA*) variables. Notably, *part* and *body* possessa are usually involved in *partitive* relationships, since a part cannot exist without a whole. Unsurprisingly, *representation* possessa are typically part of *representation* relationships, and *property* possessa most commonly appear in *feature* relationships. In addition, *activity* possessa are frequently associated with *descriptor* relationships and nominalisations (either *nom_agentive* or *nom_other*). Generally, the possessum variable is a stronger predictor for the relationship than the possessor variable.

Next, we consider all three semantic variables (*PSSR*, *PSSM*, and *RELA*) at the same time. To achieve this, we employed a novel visualisation technique called *MultiCat* (Trye et al., 2024), which is designed for exploring several categorical variables simultaneously. Researchers can analyse their own datasets in *MultiCat* by following the instructions at `https://github.com/dgt12/multicat`. Figure 8.9 is a screenshot of the *MultiCat* interface showing the most frequent combinations of semantic categories in the *Mixed Dataset*; the full dataset can be explored interactively at `https://dgt12.github.io/possession/`. Each column in the visualisation represents a different variable and each row represents a distinct combination of categories (in our case, pos-

[12] Across the whole dataset, there are 173 noun phrases that occur at least once as both possessum and possessor.

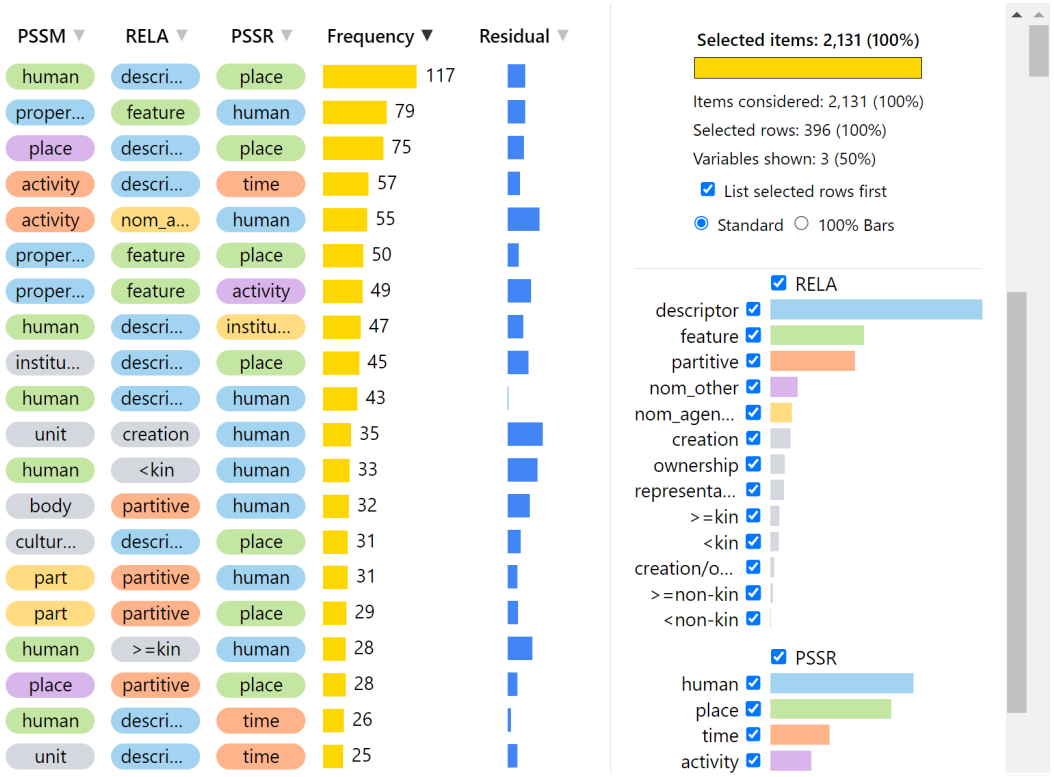


Figure 8.9: *MultiCat* visualisation of the 20 most frequent semantic category combinations for the possessum (*PSSM*), possessor (*PSSR*) and relationship (*RELA*) between the two.

sessive phrases with the same characteristics). The 'stickers' within each column are coloured according to the relative ranking of each category within the corresponding variable: blue represents the most frequent category, followed by green, orange, purple, yellow and then grey for all remaining categories. Combinations are sorted by frequency, which is visually encoded by the yellow bars. The residuals in the right-most column show the extent to which each combination is over- or under-represented in the data, using blue and red bars, respectively. The sidebar on the right side shows the marginal frequency of each category, though not all data are visible. Users can select which columns to include in the visualisation; the screenshots in this paper show only those that are relevant at each point in the analysis.

The 20 combinations shown in Figure 8.9 account for 43% of the possessive phrases observed in the *Mixed Dataset*. There are, in fact, 396 different combinations of the variables *PSSM*, *RELA* and *PSSR*, half of which occur just once. This means there is huge variation in the types of possessive phrases used in the data. Combinations that occur once or twice account for 63% of distinct configurations, while those that occur three times or fewer account for 71% of distinct configurations. The most frequent combination of *PSSM*, *RELA*, and

PSSR categories is *human-descriptor-place*, which occurs 117 times. Examples include *te Kuini o Ingarangi* 'the Queen of England' and *ngā tāngata katoa o te ao* 'all the people of the world'. Yet, while this is the single most common combination, it accounts for only 5.5% of the data overall and roughly 13% of descriptor relationships (117/882). The second most frequent combination, which occurs 79 times, involves *property-feature-human* relationships, such as *te mana o ngā tāne* 'the authority of the men'. Three of the top four combinations involve descriptor relationships, and two of these also have a *place* possessor. Interestingly, while *nom_other* relationships are more frequent than *nom_agentive* relationships, only the latter appear in Figure 8.9 (in an *activity-nom_agentive-human* relationship, e.g., *te mahi a te tangata* 'the work of the people'), suggesting that *nom_agentive* relationships exhibit less variation with respect to the possessa and possessors that they take. The residuals in the right-most column are not particularly meaningful in this figure (or any subsequent figures) since all combinations are (seemingly) over-represented. The most over-represented combination is *unit-creation-human*, due to a proliferation of tweets where the possessum is *te/ngā kōrero* and the entire possessive phrase refers to speech produced by a specific person or group of people.

Figure 8.10 shows an equivalent *MultiCat* visualisation for the *A-Only Dataset*. There are 135 combinations attested in this dataset, with Figure 8.10 accounting for 64% of the tweets where users chose *a*. Notably, the vast majority of these combinations involve *human* possessors. *Descriptor* relationships, which are predicted to take O, are far less common here than in Figure 8.9, suggesting that people do not repeatedly use *a* in the same way for these relationships (though they do still unexpectedly use *a* for this category more than any other). The most common configuration for possessive phrases with an *a* marker—accounting for one-fifth of such phrases—is *nom_agentive* relationships involving an *activity* possessum and a *human* possessor; this is also the fifth most frequent combination in Figure 8.9. The second and third most frequent configurations with *a* markers are both *creation* relationships with a *human* possessor (e.g., *ngā kōrero a Pāpā Timoti,* 'the words of Father Timothy'), which account for 11% and 5% of the data, respectively. Looking at the entire *A-Only Dataset*, *nom_agentive* and *creation* are the two most common relationships, and collectively make up half of all uses of *a*. Interestingly, all thirteen semantic relationships are represented at least once, even though most of these are associated with the O category (see Section 8.5.2).

To summarise our main findings for RQ1, while the semantic make-up of possessive phrases varies considerably, they tend to be human-centric, typically

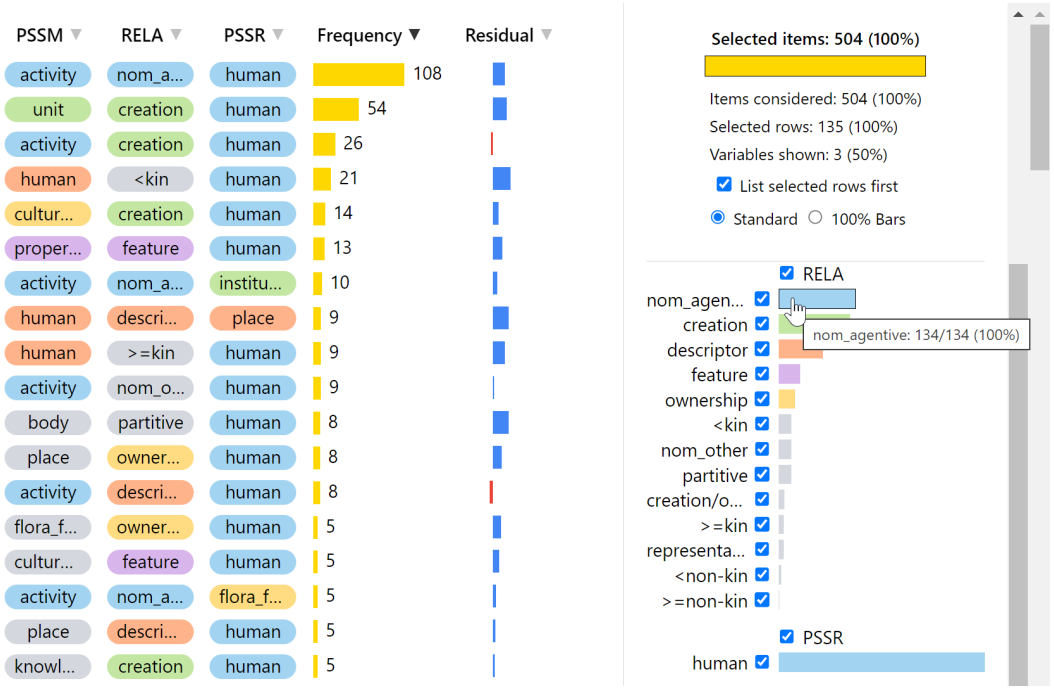


Figure 8.10: *MultiCat* visualisation of the most frequent semantic category combinations in the *A-Only Dataset*.

involving either a human possessum or possessor (though not both at the same time). *Descriptor* relationships are prevalent in the *Mixed Dataset*, and are especially likely when the possessor is a *place*. Tweeters use *a* markings most frequently for *nom_agentive* and *creation* relationships.

8.5.2 Conformity with Descriptive Rules (RQ2)

In this section, we analyse the extent to which A/O alternation in our data follows the rules described in Māori grammars, as per RQ2. We begin by considering the distribution of possessive markers in the *Mixed Dataset*; given our random sampling method, this should provide an indication of the 'real' distribution of possessive markers across different users. The dataset contains 2131 tweets, including 1980 instances of *o* (93%) and 151 instances of *a* (7%). In other words, there is a huge disparity of roughly 13 *o* markers for each *a* marker. Figure 8.11 displays a spine plot (similar to a mosaic plot) showing the composition of the *Mixed Dataset* with respect to the expected and actual markers used. The vast majority of tweets fall under the *o_expected* type (1853 tweets). More generally, tweeters use the expected marker 91% of the time (1940 out of 2131 tweets). However, this is clearly due to the extremely skewed distribution of the markers: O is the predicted category 90% of the time, and (as already mentioned) is used by tweeters 93% of the time. It

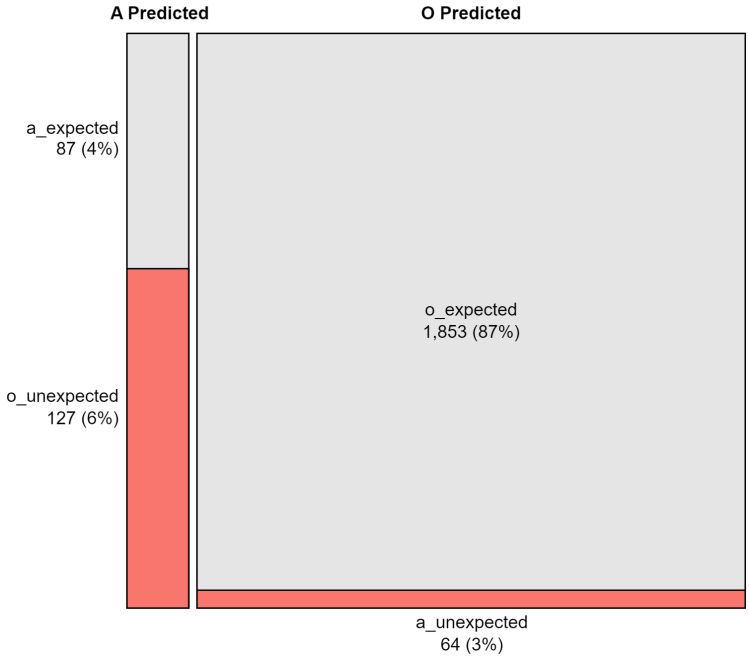


Figure 8.11: Spine plot showing the proportion of predicted a and o markers, as well as the percentage of unexpected values (red) for each marker type.

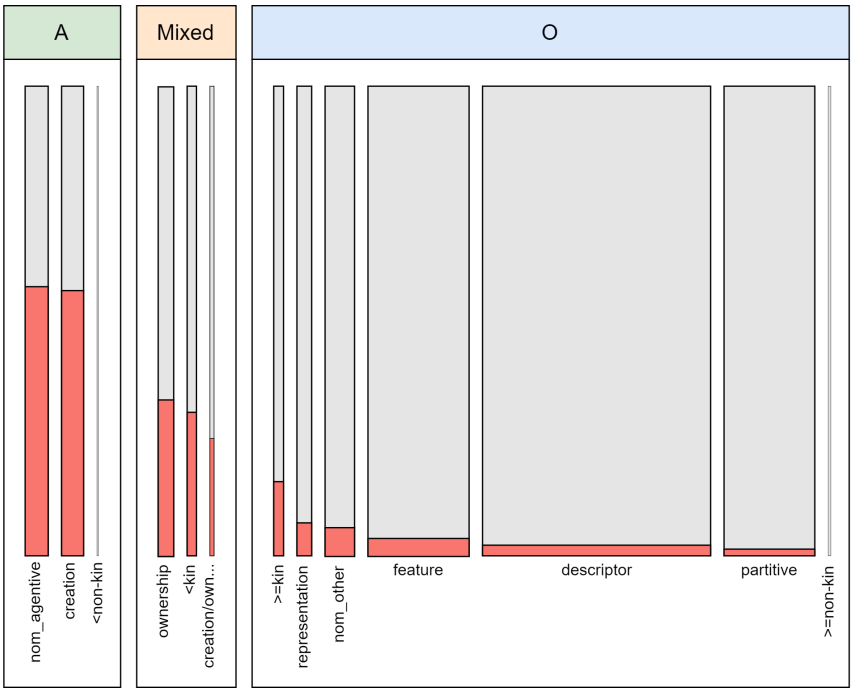


Figure 8.12: Spine plot showing semantic relationships grouped by predicted marker. Red indicates the proportion of each category with an 'unexpected' marker.

follows that tweeters use the expected marker for tweets predicted to take O virtually all of the time; this is shown by the abundance of grey in the right-hand side of Figure 8.11. In stark contrast, users adopt the expected maker for tweets predicted to take A only 41% of the time. In other words, while A-predicted tweets constitute only a small portion of the data (10%), among these tweets there are more possessive phrases (59%) where *o* is used instead of *a* (*o_unexpected*) than phrases where *a* is used as expected (*a_expected*).

Figure 8.12 displays another spine plot, this time showing the relative frequency of each semantic relationship (the area of the tiles), together with their corresponding proportions of 'unexpected' markers (the proportion of red). The categories have been grouped according to their predicted marker ('A', 'O' or 'Mixed'), as derived from Figure 8.5. We placed *<kin* in the Mixed group, rather than the A group, due to the prevalence of *uri* 'descendant' in our data, which is expected to take O rather than A. The figure shows that the relationships expected to take A tend to have the highest 'unexpected' rates. The *nom_agentive* and *creation* relationships are, in fact, the only two categories for which people use the 'unexpected' marker more often than the 'expected' one (with 'unexpected' rates of 57% and 56%, respectively). Conversely, relationships that are expected to take (only) O have very low 'unexpected' rates, with none exceeding 16% (*>kin*) and all others being less than half of this value. The Mixed categories fall somewhere in between; unsurprisingly, most (75-82%) of their 'unexpected' values come from possessive phrases in which an *a* marker was predicted but an *o* marker was used instead. Overall, ignoring the very infrequent *non-kin* categories, users were most likely to employ the expected marker for *partitive* and *descriptor* relationships. These findings suggest that people are increasingly using O forms in situations where grammars specify A forms.

Given that most discrepancies between the usage described in grammars and observed in our data occur when an *o* marker was unexpectedly used instead of an *a* marker (*o_unexpected*), we now focus on these specific cases. There are 127 such possessive phrases, which we acknowledge is a very small sample size. In line with the previous chart, Figure 8.13 shows that *nom_agentive* relationships attract the largest number of unexpected markers, although *creation* relationships are a close second. The prevalence of unexpected markers for *nom-agentive* relationships suggests that speakers may prioritise semantic information (i.e., the kinds of entities involved in a possessive relationship) over syntactic criteria (i.e., the type of verb). The most frequent combinations in which an *o* marker is used instead of an *a* marker are *activity-*

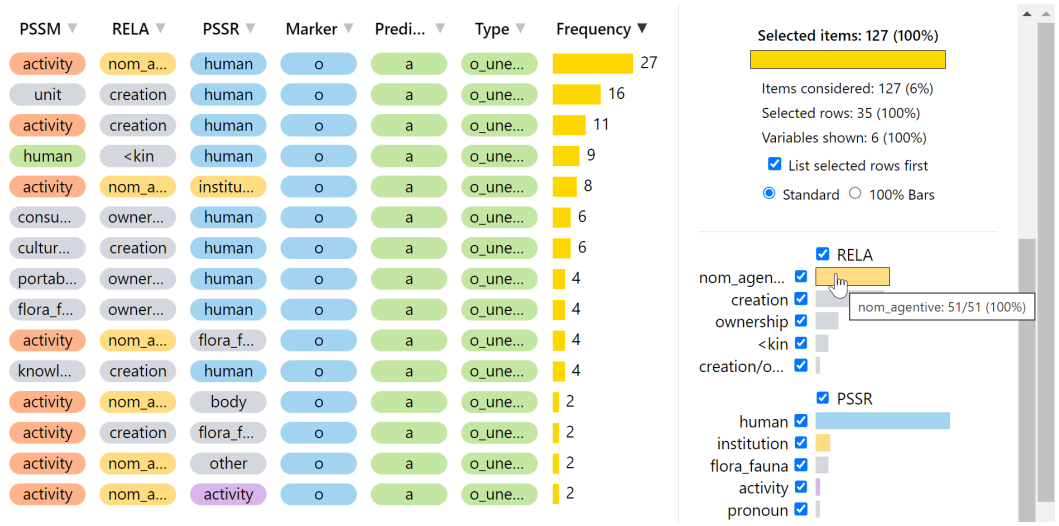


Figure 8.13: Recurrent configurations in which an *o* marker was used instead of an *a* marker.

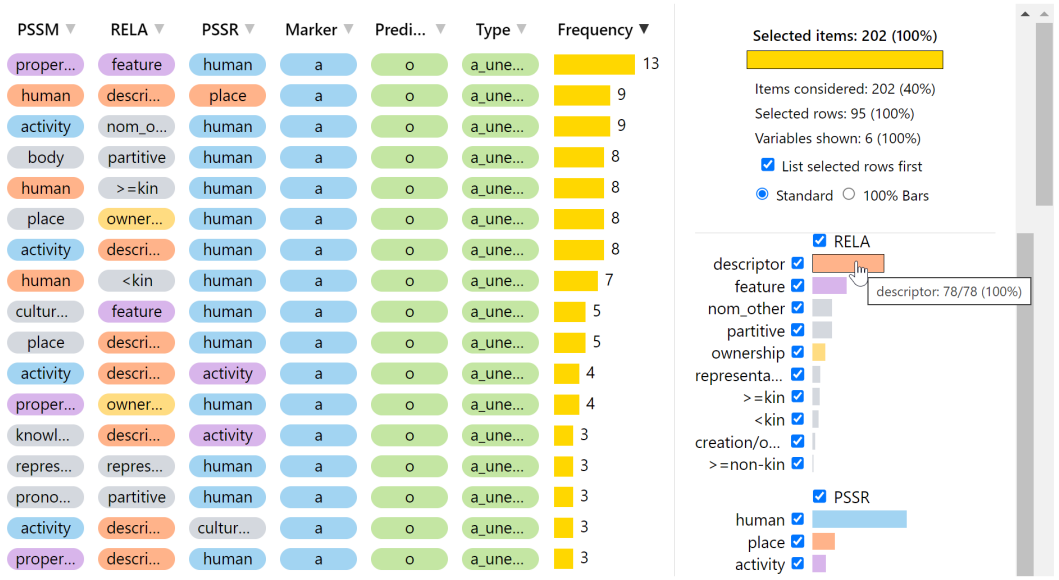


Figure 8.14: The most common configurations in the *A-Only Dataset* in which an *a* marker was used instead of an *o* marker.

nom_agentive-human relationship (27 instances; e.g., *ngā manaakitanga *o te wāhi* 'the generosity of the place'), followed by a *unit-creation-human* relationship (16 instances, all of which have a possessum to do with 'words'). Nearly three-quarters of the *o_unexpected* phrases involve a human possessor (74%), and just under half use an *activity* possessum (48%), most of which are linked to a *nom_agentive* relationship.

One anonymous reviewer asked whether the patterns observed could be attributed to our characterisation of *activity* possessa. Looking at every tweet with an *activity* possessum across the *Mixed Dataset*, 69% of the data were

predicted to take an *o* marker, and *o* was used as expected 96% of the time. However, among the tweets with an *activity* possessum predicted to take *a*, the *a* marker was used only 42% of the time. This suggests that tweets involving activities are indeed contributing to (possible) language change in progress regarding the use of the A/O categories.

Figure 8.14 shows the opposite kind of unexpected markers to Figure 8.13, whereby an *a* form was unexpectedly used instead of an *o* one, this time homing in on the *A-Only Dataset*. This applies to 40% of tweets in this dataset (n=202); the other 60% (n=302) are tweets for which an *a* marker was expected. The most common use of *a* instead of *o* is for *property-feature-human* relationships (13 occurrences; e.g., *te riri *a Tāwhirimātea* 'the anger of Tāwhirimātea [the god of weather]'). The next most common relationships, with 9 occurrences each, are *human-descriptor-place* and *activity-nom_other-human*. The most unexpected uses of *a* occur with *activity, property*, or *human* possessa, which are also the three most frequent categories for which *o* is used, and, as with the inverse kind of unexpected marker, often with a *human* possessor (53%). Overall, *descriptor* relationships have the most unexpected *a* markers, followed by *feature, nom_other*, and *partitive* relationships.

Since we have timestamps for all the tweets in our data, we can check whether there are any trends in usage of the *a/o* markers across time. Figure 8.15 summarises the use of each form in two panels: *o* on the left and *a* on the right. These are expressed as percentage splits between the expected form (in blue), which adheres to grammatical descriptions, and the unexpected form (in red), which deviates from grammatical descriptions. For example, for 2022, we can see that *all* uses of *o* were marked by the (expected) *o* form, whereas 80% of expected uses of *a* were instead coded by *o*. In general, the *o*-possessives are marked according to the descriptions given in grammars, and increasingly so over time (the amount of red gradually decreases in the left-hand panel), whereas *a*-possessives appear to be increasingly marked by *o*. The trend lines in each chart corroborate these findings, although the relationship is weak in both cases, as indicated by the R^2 values of 0.281 and 0.253, respectively.

Overall, it would appear that Māori-language tweeters use *o* in most contexts, regardless of the inherent semantic relationship between the possessum and possessor, with users adopting the expected marker for tweets predicted to take *a* less than half of the time (41%) according to our *Mixed Dataset*. These findings support our hypothesis for RQ2.

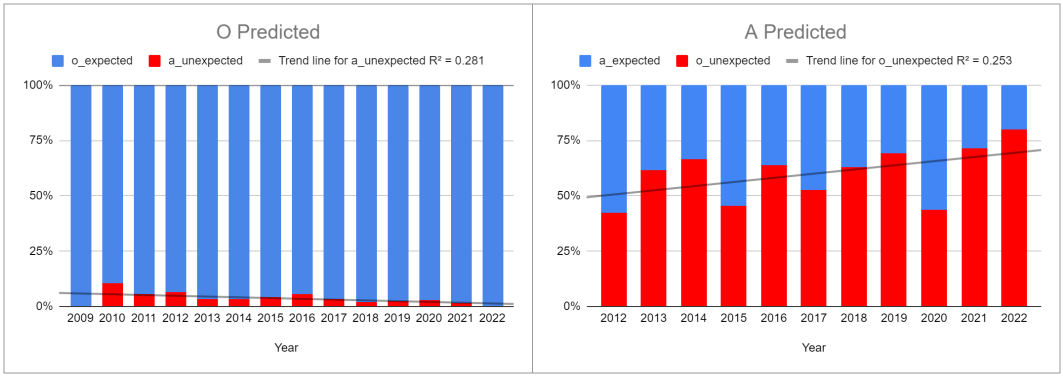


Figure 8.15: 100% stacked bar charts showing the proportions of expected and unexpected forms for each marker in the *Mixed Dataset* across time.

8.5.3 Sociolinguistic Characteristics of Tweeters (RQ3)

Our third and final research question relates to the social characteristics of Māori-language tweeters and whether there are any trends concerning their usage of *a/o* alternation. As a first step, it is useful to explore the macro-level patterns concerning all the available demographic variables for the users in our *Mixed Dataset*. Figure 8.16 provides such an overview. This visualisation, derived from the *Heatmap Matrix Explorer* (Trye et al., 2023; cf. Rocha and da Silva, 2018), aims to provide a compact summary of a large collection of categorical variables, with darker values indicating higher counts of pairwise category intersections.

The following variables are included in the heatmap: *Gender, Location* (New Zealand vs overseas), *Ethnicity, Years Active* (the number of distinct years for which the user has at least one tweet in the RMT Corpus), *% Reo Content* (Low: <40%, High: >=40%) and *Followers* (<1000, >=1000). Unfortunately, not all tweeters' location, ethnicity or percentage of Māori-language tweets were available, as indicated by the presence of the "Unknown" category. The exact age of each tweeter was not available, but we suspect that most users in the corpus fall within 25–55 years of age

The two left-most columns in Figure 8.16 reveal a balanced distribution between male and female tweeters, both overall and with respect to each pairwise intersection. In other words, the number of male tweeters from each location, ethnicity, time interval, etc. is similar to the corresponding number of female tweeters. Similarly, there is a relatively even split between the number of tweeters with fewer and more than 1000 followers, though tweeters whose contributions span fewer than six distinct years in the corpus tend to have fewer followers.

Figure 8.16: *Heatmap Matrix* visualisation showing information about tweeters in the *Mixed Dataset*.

It is evident from Figure 8.16 that most tweeters are based in New Zealand and of Māori descent, with only a small proportion of known Pākehā tweeters (roughly 6%). The location of these tweeters is expected, as New Zealand is the only place in the world where Māori is an official language (migrants who speak Māori can be found elsewhere, but only five such tweeters were identified in the *Mixed Dataset*). However, while most Māori-language tweeters are ethnically Māori, it is interesting to note that some Pākehā are also committed to supporting the language by using it on social media. The majority of tweeters contributed data to the RMT Corpus over a period from six to ten distinct years, demonstrating a sustained commitment to Māori-language tweeting. Notably, however, only 14 users tweet in Māori more than 40% of the time, and among them, only three have more than 1000 followers. Tweeters in the *Low % Reo Content* category may still tweet in Māori on a regular basis, just not as often as they post in English or other languages.

We now focus specifically on *Gender*, since we consider it to be the most interesting variable for which all values are known. Figure 8.17 compares the

proportions of unexpected *a* and *o* markers produced by males and by females. Contrary to our hypothesis, males produced slightly higher rates for both kinds of unexpected marker. Breaking this down further by each relationship (Figure 8.18), males produced larger proportions of unexpected markers for most relationships. A possible explanation for this is that women are more likely than men to learn Māori through formal education (Te Kupenga, 2020), where the importance of the A/O categories is often emphasised. However, the differences are not large enough to be deemed statistically significant, which means we cannot confidently conclude that males are less likely to use an expected marker than females (or vice versa).

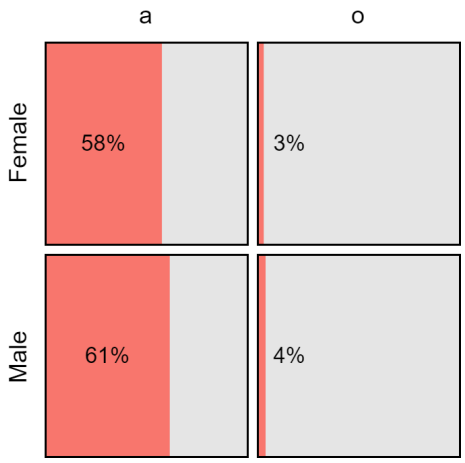


Figure 8.17: Proportions of 'unexpected' markers (red) for each predicted marker type, broken down by gender.

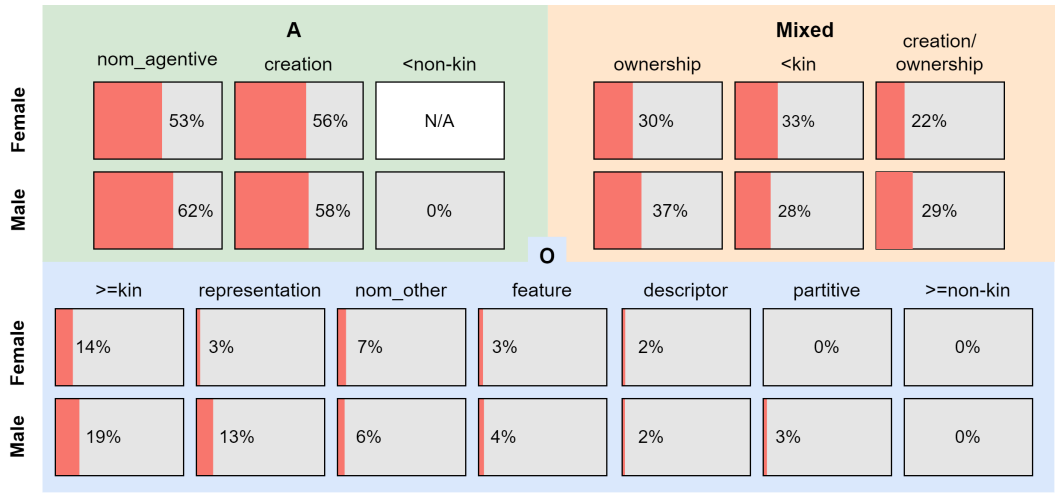


Figure 8.18: Proportions of 'unexpected' markers (red) for each relationship, broken down by gender. The same bin size is used for ease of comparison.

8.6 Discussion & Conclusion

We begin this section with an overview of our findings. Using Twitter data and a combination of manual and automatic methods, we extracted and analysed A/O possessives of the form [possessum *a/o* possessor], such as *te mahi a te tangata* 'the work of the people' and *ngā rangatira o Waikato* 'the chiefs of Waikato'. Our analysis uncovered a wide array of possessive types and tokens: users employ this construction to express possessa encoding humans, places, activities, properties, parts (of entities) and units, as 'possessed'—either literally or metaphorically—by humans, places, time, institutions and activities, among others (Figure 8.6). In other words, we find that tweeters make full use of the varied semantic categories and a wide range of noun phrases. What may be a slight surprise in the range of categories expressed is the comparatively low frequency of kinship relationships mentioned. Given the chief importance that family (*whānau*) and genealogy (*whakapapa*) play in Māori culture, it may be unexpected to find so few occurrences of kinship in these data. However, this may be because we considered only a subset of the possessive domain, focusing on just one of several possible constructions. Kinship relationships are perhaps more likely to occur in other forms of possession, such as with possessive pronouns (*tōu māmā* 'your mother'; *ō māua mātua* 'our parents').

A second aspect of interest, and, for some, of great concern given some of the negative publicity of language used on social media (see discussion in Bleaman, 2020), relates to the patterns of A/O alternation and their divergence from existing grammars and learner texts. With this being the first quantitative study (to our knowledge) of naturally occurring possessives in Māori, our findings confirm the frequently noted trend that O constitutes the unmarked category and A the marked one (Clark, 1976, pp. 42–44; Bauer et al., 1997, p. 391; Harlow, 2007, p. 168). Moreover, we find that users nearly always match descriptions in Māori grammars with respect to the use of *o* (97%), and where differences arise, these mostly concern the use of *a* (Figures 8.11 and 8.12). In total, 42% of users in the *Mixed Dataset* do not use *a* at all, which is significant. Peering more deeply at constructions that would be A-marked according to grammars, we find that differences in use (i.e., the use of *o* instead of an expected *a*) are most frequent with nominalisations; in other words, in cases where deciding on the appropriate marker involves grammatical—rather than semantic—criteria.

There are several ways to explain these patterns. First, they are in line with usage-based theories (Barlow and Kemmer, 2000) and exemplar theory (Pierrehumbert, 2001). Speakers execute 'best' what they see and hear most:

the O category is very frequent, and speakers' perception of this use is further reinforced through recurrent use. Conversely, A is infrequent, and uncertainty around its use leads to further mismatches between expectations coded in the grammar of the language and learner texts on the one hand, and actual use on the other. Another way to interpret this pattern is by recourse to language change: the data reflect a change in progress from an alternation to a shift towards one form. The two interpretations are not in contradiction; one can, in principle, lead to the other, and there is some support for this (Figure 8.15), but further data are needed to confirm this at a macro-level. Grammatical levelling towards the O category in the possessive system has already been asserted by Baclawski (2011), but without recourse to substantial data. We remain cautious about this claim, because our data need further validation from additional genres. Moreover, given the relationship between A/O possessives and the conceptualisation of control and agency in various relationships, it may well be the case that what we ultimately end up seeing is not a complete levelling in the system, but a slight reorganisation of it. This reorganisation could be based around the frequency effects (Haspelmath, 2021) of certain noun phrases that are strongly associated with the A category, and, hence, perpetuate its use.

We hope that our findings will be helpful for teachers and learners of Māori, who may benefit from a flowchart approach to learning the A/O categories, inspired by the algorithm in Figure 8.5. This study aims to bring attention to lesser-known aspects of the possessive system, such as the different types of relationship that can occur, especially *nom_agentive* and *creation* relationships, for which *o* was unexpectedly used more often than *a*.

Looking beyond these conclusions, we hope to have demonstrated the benefits of analysing a social media corpus of a low-resourced language, and of using visualisation tools such as *MultiCat* and the *Heatmap Matrix Explorer* to analyse categorical data. The scarcity and limitations of existing Māori corpora are particularly constraining for studies of Māori syntax, which require large-scale data. In this context, social media afford useful resources that accelerate academic research, albeit with certain tradeoffs: tweets come packaged in short messages and are notoriously noisy.

One benefit of social media content, apart from its relative availability and size, is the opportunity to tap into the language produced by younger speakers (Keegan and Cunliffe, 2014). For indigenous languages like Māori, this presents an opportunity to analyse language use among a socio-demographically distinct group of individuals, who are likely to be younger speakers with access

to digital tools and higher education, and who may be active members in language revitalisation efforts. The focus on younger speakers is particularly beneficial for the study of Māori, because, according to current research, the generation born between 1984 and 2003 accounts for almost half of more proficient speakers (Lane, 2024). We note, however, that Twitter/X appears to be less popular among adolescents in Aotearoa than platforms such as Instagram, Snapchat, and TikTok (Goodwin et al., 2024). Indeed, Māori speakers under the age of twenty may be driving language change on these platforms.

Our analysis has several limitations, but also provides rich opportunities for future work. We may have unwittingly removed relevant tweets through our data collection process and acknowledge that our data may not capture general population norms. Additional data are needed to perform complex modelling in order to identify sociolinguistic patterns of change. Furthermore, analyses of semantic content on Twitter are difficult, as tweets may lack sufficient context when read in isolation. Classifying semantic relationships inevitably comes with potential for disagreement. Finally, our algorithm for predicting the expected marker may have introduced errors, especially in cases where either marker could be seen to conform with grammatical descriptions, but only one was treated as expected. Future work could involve exploring additional variables that may be relevant, such as number and specificity, extending our analysis to the many other possessive constructions in Māori that involve A/O alternation, and analysing historical corpora to better understand the evolution of these categories over time. In addition, it is worth investigating whether the A/O categories are sensitive to sociolinguistic effects, with A denoting possessors of a (perceived) higher status, in line with Thornton (1998).

We leave the reader with a final thought: if A/O alternation in Māori were to disappear completely, what exactly would be lost? Given that this contrast, at its core, steers speakers towards considering subtle aspects of the relationship between the possessum and possessor, about agency, ownership, and responsibility, its preservation provides a window into the Māori worldview, shedding light on how Māori understand their relationship with the physical environment (Kārena-Holmes, 2021), with their *whakapapa*, *whānau*, and community, and with the objects and entities around them. It is astounding to consider how much can be packed into such a tiny part of the linguistic system. Indeed, small words have significant implications: *He iti te kupu, he nui te kōrero.*

Acknowledgements

The authors thank the three anonymous reviewers for their insightful comments and suggestions. We gratefully acknowledge Rangatahi Tahere for his help translating the data. DT and AC thank Beau Stowers for sparking their interest in this topic back in 2019. The online tool `https://app.diagrams.net` was used to create several figures in the paper.

8.7 Postscript

To the best of our knowledge, this chapter has presented the first quantitative study of possession in Māori, as well as the first linguistic study of Māori based on social media data. Our analysis of the [possessum a/o possessor] construction indicates that the O category is overwhelmingly dominant, occurring in 93% of the inspected tweets, and that it is used more often than not (59% of the time) in cases where Māori grammars would designate the A category. More generally, this chapter has shown how a social media corpus of a low-resourced and endangered language can be analysed to learn more about that language, providing fresh insights into the linguistic practices of its speakers.

Furthermore, we have demonstrated how the categorical visualisation techniques discussed in Part II of this thesis can be used to increase understanding of a specific application domain: both MultiCat and the Heatmap Matrix Explorer helped to reveal information that was not apparent from the raw data alone. MultiCat was used to examine the most commonly recurring semantic patterns, including those related to unexpected uses of each possessive marker. Although not mentioned in the paper, MultiCat was also helpful for identifying mistakes in data annotations, such as a handful of tweets whose 'Type' was inconsistent with the relationship given; these tweets were subsequently corrected. The Heatmap Matrix, on the other hand, provided a high-level overview of the different types of tweeters in the corpus. Interacting with these visualisations also provided inspiration for other figures and statistics, such as the spine plots, which combine elements from successive MultiCat queries into a single figure with an appropriate level of detail.

8.8 References

Aikhenvald, A. Y. (2013). Possession and ownership: A cross-linguistic perspective. In Aikhenvald, A. and Dixon, B., editors, *Possession and Ownership: A Cross-Linguistic Typology*, pages 1–64. Oxford University Press.

Baclawski, K. (2011). A/O possession in modern Māori. Unpublished manuscript, Dartmouth College. Available online: `https://linguistics.berkeley.edu/~kbaclawski/Baclawski_2011_Maori_possession`.

Barlow, M. and Kemmer, S. (2000). *Usage-Based Models*. CSLI Publications.

Bauer, W., Parker, W., and Evans, T. (1993). *Maori*. Routledge.

Bauer, W., Parker, W., Evans, T., and Teepa, T. (1997). *The Reed Reference Grammar of Māori*. Reed.

Biggs, B. (1955). The compound possessives in Maori. *The Journal of the Polynesian Society*, 64(3):341–348.

Biggs, B. (1996). *Let's Learn Maori: A Guide to the Study of the Maori Language*. Auckland University Press.

Biggs, B. (2000). Te paanui a wai-whararikī. Māori Department, University of Auckland.

Bleaman, I. L. (2020). Implicit standardization in a minority language community: A real-time syntactic change among Hasidic Yiddish writers. *Frontiers in Artificial Intelligence*, 3:1–20.

Capell, A. (1949). The concept of ownership in the languages of Australia and the Pacific. *Southwestern Journal of Anthropology*, 5(3):169–189.

Chousou-Polydouri, N., Inman, D., Huber, T. C., and Bickel, B. (2023). Multivariate coding for possession: Methodology and preliminary results. *Linguistics*, 61(6):1365—1402.

Christensen, I. S. (2003). Proficiency, use and transmission: Maori language revitalisation. In *New Zealand Studies in Applied Linguistics*, volume 9, pages 41–61.

Clark, R. (1976). *Aspects of Proto-Polynesian Syntax*. Te Reo Monograph. Linguistic Society of New Zealand, Auckland, N.Z.

Foster, J. (1987). *He Whakamārama: A New Course in Māori*. Heinemann.

Fusi, V. (1985). Action and possession in Māori language and culture: A Whorfian approach. *L'Homme*, pages 117–145.

Goodwin, I., Lyons, A. C., Young, J., and Neha, T. (2024). Young people's internet use, social media activity, and engagement with social media influencers. `https://researchspace.auckland.ac.nz/handle/2292/68247`.

Greensill, H., Manuirirangi, H., and Whaanga, H. (2017). Māori language resources and Māori initiatives for teaching and learning te reo Māori.

In Whaanga, H., Keegan, T. T. A. G., and Apperley, M., editors, *He Whare Hangarau Māori - Language, culture & technology*, pages 1–9. Te Pua Wānanga ki te Ao / Faculty of Māori and Indigenous Studies, the University of Waikato, Hamilton, New Zealand.

Harlow, R. (2000). Possessive markers in Māori. *STUF-Language Typology and Universals*, 53(3-4):357–370.

Harlow, R. (2007). *Maori: A Linguistic Introduction*. Cambridge University Press.

Harlow, R. (2015). *A Māori Reference Grammar (2nd ed.)*. Huia Publishers.

Harlow, R., Bauer, W., Maclagan, M., Watson, C., Keegan, P., and King, J. (2011). Interrupted transmission and rule loss in Māori: The case of ka. *Oceanic Linguistics*, pages 50–64.

Haspelmath, M. (2017). Explaining alienability contrasts in adpossessive constructions: Predictability vs. iconicity. *Zeitschrift für Sprachwissenschaft*, 36(2):193—231.

Haspelmath, M. (2021). Explaining grammatical coding asymmetries: Form–frequency correspondences and predictability. *Journal of Linguistics*, 57(3):605—633.

Head, L. (1989). Making Maori sentences. Longman Paul. Available online: `https://tereomaori.tki.org.nz/content/download/2780/15817/file/moe626-making-sentences-complete-100dpi.pdf`.

Higgins, R., Rewi, P., and Olsen-Reeder, V., editors (2014). *The Value of the Māori Language: Te Hua o te Reo Māori*, volume 2. Huia Publishers.

Kārena-Holmes, D. (2021). *Te Reo Māori: The Basics Explained*. Oratia Media Ltd.

Keegan, T. T. A. G. and Cunliffe, D. (2014). Young people, technology and the future of te reo Māori. In Higgins, R., Rewi, P., and Olsen-Reeder, V., editors, *The Value of the Māori Language: Te Hua o te Reo Māori*, pages 385—-398. Huia Publishers.

Kelly, K. (2014). Iti te kupu, nui te kōrero - The study of the little details that make the Māori language Māori. In Higgins, R., Rewi, P., and Olsen-Reeder, V., editors, *The Value of the Māori Language: Te Hua o Te Reo Māori*, pages 255–267. Huia Publishers.

Kelly, K. (2015). *Aspects of change in the syntax of Māori - A corpus-based study*. Doctoral thesis, Te Herenga Waka-Victoria University of Wellington.

King, J. (2018). Māori: Revitalization of an endangered language. In Rehg, K. L. and Campbell, L., editors, *The Oxford Handbook of Endangered Languages*, pages 592–612. Oxford University Press.

Krupa, V. (1964). On the category of possession in Maori. *Bulletin of the School of Oriental and African Studies*, 27(2):433–435.

Krupa, V. (2003). Extralinguistic basis of the category of possessivity. *Asian and African Studies*, 12(2):122–134.

Lane, C. (2024). First and second language speakers in the revitalisation of te reo Māori: A statistical analysis from Te Kupenga 2018. *Te Reo*, 66(2):28–56.

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.

Moorfield, J. C. (1988). *Te Kākano*. Longman Paul, Auckland.

Nicholas, S. (2010). An investigation of the so-called 'passive' construction in New Zealand Māori. Master's thesis, The University of Auckland.

Pierrehumbert, J. B. (2001). Exemplar dynamics, word frequency, lenition, and contrast. In Bybee, J. and Hopper, P., editors, *Frequency effects and the emergence of linguistic structure*, pages 135–157. John Benjamins.

Ryan, P. M. (1974). *The New Dictionary of Modern Māori*. Heinemann.

Scannell, K. P. (2022). Managing data from social media: The Indigenous Tweets project. In Berez-Kroeker, A. L., McDonnell, B., Koller, E., and Collister, L. B., editors, *The Open Handbook of Linguistic Data Management*. MIT Press.

Statistics NZ (2018). 2018 census totals by topic: National highlights updated.

Stefanowitsch, A. (2020). *Corpus linguistics: A guide to the methodology*. Language Science Press.

Tawhara, T. A. M. (2015). Kia Māori te reo Māori? An investigation of adult learner attitudes towards the impact of English on te reo Māori. Master's thesis, University of Otago.

Te Kupenga (2020). More than 1 in 6 Māori people speak te reo Māori. `https://www.stats.govt.nz/news/more-than-1-in-6-maori-people-speak-te-reo-maori`.

Thornton, A. (1998). Do a and o categories of "possession" in Maori express degrees of tapu? *The Journal of the Polynesian Society*, 107(4):381–393.

Trye, D., Apperley, M., and Bainbridge, D. (2023). Extending the Heatmap Matrix: Pairwise analysis of multivariate categorical data. In *2023 27th International Conference Information Visualisation (IV)*, pages 29–36. IEEE.

Trye, D., Apperley, M., and Bainbridge, D. (2024). Multicat: A visualisation technique for multidimensional categorical data. [Unpublished Manuscript].

Trye, D., Keegan, T. T., Mato, P., and Apperley, M. (2022). Harnessing Indigenous Tweets: The Reo Māori Twitter corpus. *Lang Resources & Evaluation*,

56:1229–1268.

Whaanga, H. and Greensill, H. (2014). An account of the evolution of language description of te reo Maori since first contact. In Onysko, A., Degani, M., and King, J., editors, *He Hiringa, He Pūmanawa: Studies on the Māori language*, pages 7–32. Huia.

Williams, H. W. and Williams, W. L. (1971). *First Lessons in Māori*. Government Printer.

Wilson, W. H. (1982). Proto-Polynesian possessive marking.

Chapter 9

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

This chapter presents a case study on the co-occurrence of Māori loanwords in New Zealand English newspaper articles. Accordingly, we shift our attention from conversational Twitter data—which was the focus of Chapters 6–8—to the more formal, carefully edited language found in mainstream media. While the previous chapter, and all of Part II, dealt with what might be called 'typical' categorical data, because it did not have any inherent additional structures, in this chapter we consider *relational data* (networks) with multiple categorical attributes. It was noted in Chapter 3 that categorical variables can occur in datasets with special properties, such as time-oriented and geospatial data, and relational data is no exception.

In order to investigate which Māori loanword types appear within the same article, we model our data using both networks and hypergraphs. The difference between these two structures is that the edges in a network connect exactly two nodes, while those in a hypergraph can connect any number of nodes, enabling, in our case, the preservation of the entire set of loans in each article. The nodes/elements in our analysis represent distinct loanwords, whereas edges represent loanwords that co-occur within the same article. A range of different visualisation techniques exist for representing such data; however, we employ standard node-link diagrams and a recent technique called PAOHVis

(Valdivia et al., 2019) to visualise the networks and hypergraphs, respectively.

Importantly, each loanword in our data is tagged according to three categorical variables: semantic domain, size, and listedness. We use these variables to explore whether similar groupings of loanwords occur across different articles. Contrary to the approach used in other chapters, we consider the effect of each categorical variable *in turn* (via the colour and shape of each node), rather than analysing them all together. To extract high-level patterns, we *aggregate* the hypergraphs by adapting the PAOHVis technique. These modifications, which can theoretically be applied to any PAOHVis representation with categorical attributes, were presented in a poster paper, reproduced in Appendix G (pp. 349–352).

Publication Details

The following paper has been reproduced with minor changes to the formatting, as discussed in Section 1.4:

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*, *28*(4), 461–499. https://doi.org/10.1075/ijcl.21124.try

Abstract

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.

9.1 Introduction

As we live in an ever more 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: English. This particular variety is called New Zealand English (hereafter, NZE; see Section 9.2.2).

Given the large body of work focusing on loanwords, several 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 9.3) 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. 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

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

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 9.4.3) 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?

9.2 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 9.2.1). Additional background information about the language contact situation in New Zealand is then given, together with a summary of related work (Section 9.2.2), as this provides necessary context for understanding our case-study.

9.2.1 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; Calude et al., 2020b). In such studies, the loanwords of interest are typically examined independently, without considering 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 9.3.1, 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, 2010; 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; Calude et al., 2019), such that Māori-related topics elicit higher counts.

Another factor relevant to entrenchment is acceptance or listedness (see Section 9.3.4). Discussing code-mixing, Muysken (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 and Deuchar (2012:631) for English verbs borrowed into Welsh, by verifying the listedness of these verbs in Welsh dictionaries. According to Stammers and

Deuchar (2012:631), such entries show loanwords that are 'established borrowings'.

9.2.2 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, 2006:18). 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 (2006), 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 Calude et al., 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; Levendis & Calude, 2019; Trye et al., 2020) or their relative success (Calude et al., 2020b). 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.

9.3 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 is available on our companion website (Kiwi Words, 2021). Section 9.3.1 details the corpus used, and Section 9.3.2 explains our criteria for selecting loanwords. We then explain how we computed loanword co-occurrence (Section 9.3.3), outline three linguistic properties of interest (Section 9.3.4), and provide an overview of the loanwords' overall frequency in the corpus (Section 9.3.5).

9.3.1 Overview of the Matariki Corpus

This study investigates loanword co-occurrence within an existing corpus of NZE newspaper articles, called the Matariki Corpus (Calude et al., 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 9.1. The corpus has a high loanword rate, likely because the topic of discourse is directly relevant to Māori.

Table 9.1: 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

9.3.2 Loanword selection process

The method used for identifying loanwords in the Matariki Corpus involved a combination of computational and manual techniques. There were four key steps, as detailed in Figure 9.1.

The first step was to identify potential Māori loanwords in the corpus, by leveraging code developed by Te Hiku Media (2019). The *kupu_tūtira* script was used to obtain a list of words that are consistent with Māori orthography.

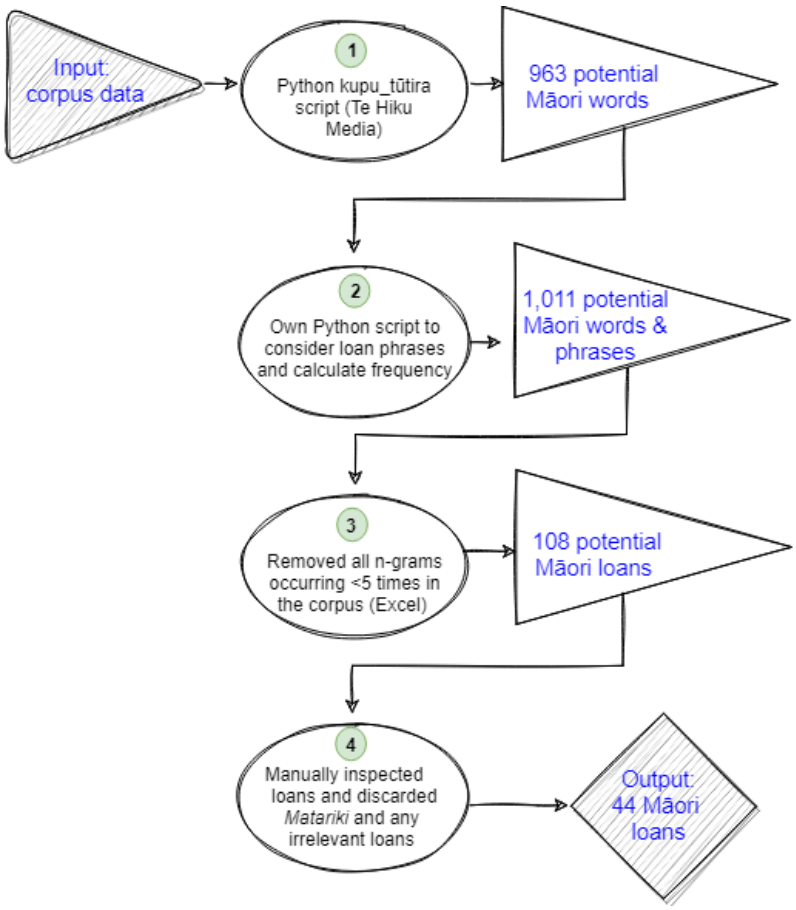


Figure 9.1: The loanword selection process.

Such items: (i) consist solely of characters in the Māori alphabet;[2] (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 to 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 9.2.1), 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

[2]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)*.

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 9.3.3). For instance, the script incorrectly extracted *hope more maori* as a nonce loan-phrase of size three, when the first two words were 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).

In step four, we manually inspected the remaining loans to remove false positives: (i) non-loans that happened to conform with Māori orthography (e.g. *make*) and (ii) most proper nouns referring to personal names (e.g. *Hone Pene*) and places (e.g. *Rotorua*), unless they had an 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 original 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 H for details).

9.3.3 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 9.1). 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 9.5. 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 9.4.2) and hypergraphs (Section 9.4.3). 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.

9.3.4 Linguistic properties

Following previous work, we identified three linguistic properties that are relevant to Māori loanword use in NZE (see Section 9.2.2), namely semantic domain, size, and listedness. We coded our set of 44 loans with respect to each of these variables; see Appendix H.

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. *pōwhiri* "welcoming ceremony"). The last two categories are not always straightforward to disambiguate; the crucial difference between them is that the former constitute tangible objects, 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; Calude et al., 2020a) 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 Calude et al., 2020b). 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 Muysken (2000), and operationalised according to Stammers and Deuchar (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 9.2 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

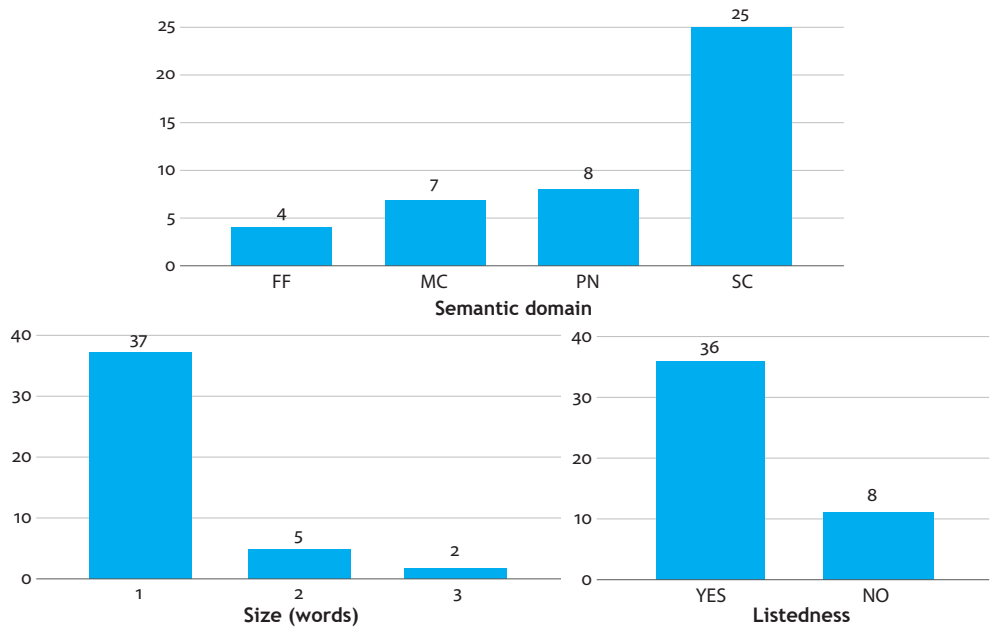


Figure 9.2: Linguistic properties of the 44 loan types of interest (in the top chart, 'FF'=flora and fauna, 'MC'=material culture, 'PN'=proper noun and 'SC'=social culture).

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

We now contrast the number of types per category with the number of tokens per category, summarised in Figure 9.3. Since *Māori* is an outlier in the corpus (see Figure 9.4), we display counts both with and without this loan, using black and blue bars, respectively. This distinction is important because we use the data represented by the black bars for the network analysis (Section 9.4.2) and the data represented by the blue bars for the hypergraph analysis (Section 9.4.3). 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.

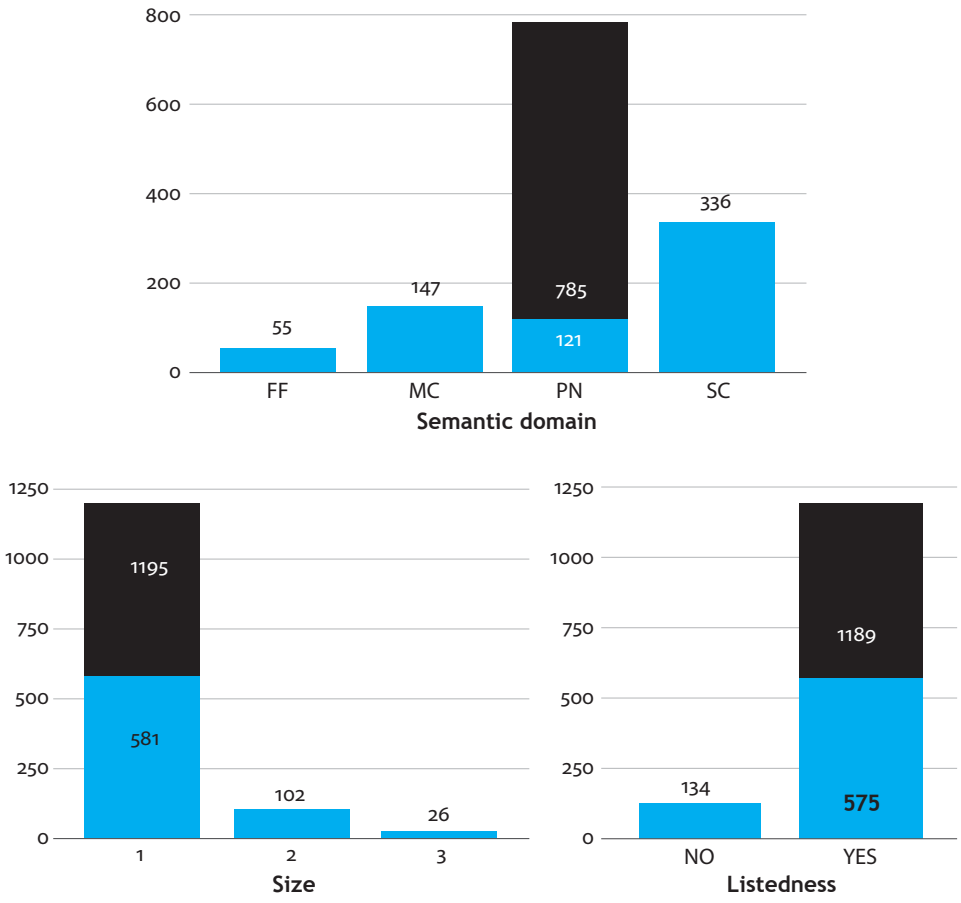


Figure 9.3: Linguistic properties aggregated by number of tokens per category.

9.3.5 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=41,[3] 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 9.4 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.[4]

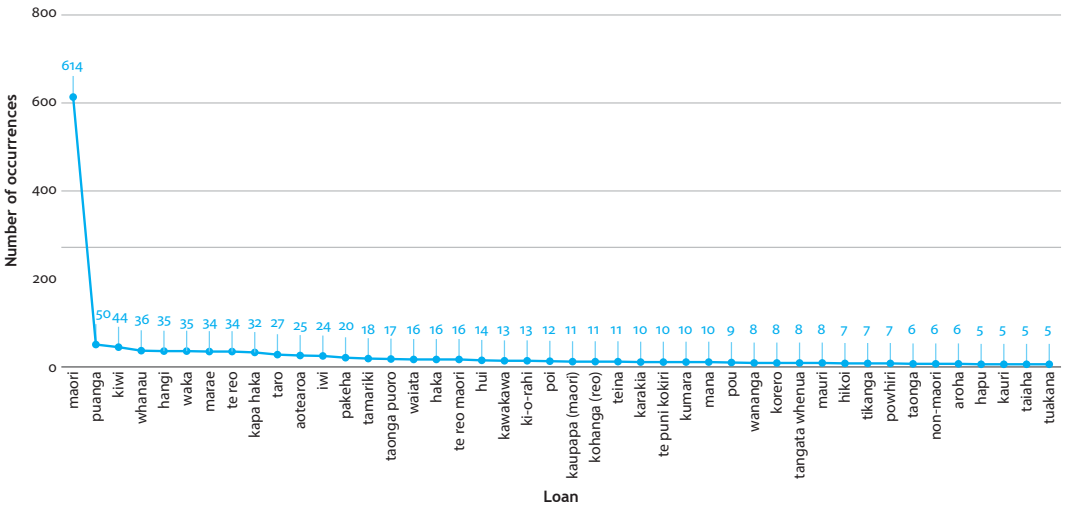


Figure 9.4: Raw frequency of productive loans in the corpus.

9.4 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 9.4.1), then present standard networks, in which the loan sets are 'flattened' into pairwise co-occurrences (Section 9.4.2). Finally, we take a closer look at the loan sets as

[3]The degrees of freedom used reflects the removal of the loan *Māori* from the data, which is an outlier (see Figure 9.4).

[4]This is also the case for Figures 9.6–9.11.

a whole, using a more robust representation, albeit less widely used in language analyses, called a 'hypergraph' (Section 9.4.3).

9.4.1 Distribution of loan types

We can study co-occurrence relationships by looking at the number of loans in each text. Figure 9.5 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 9.5, 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 pink bars on the left-hand panel of Figure 9.5 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.

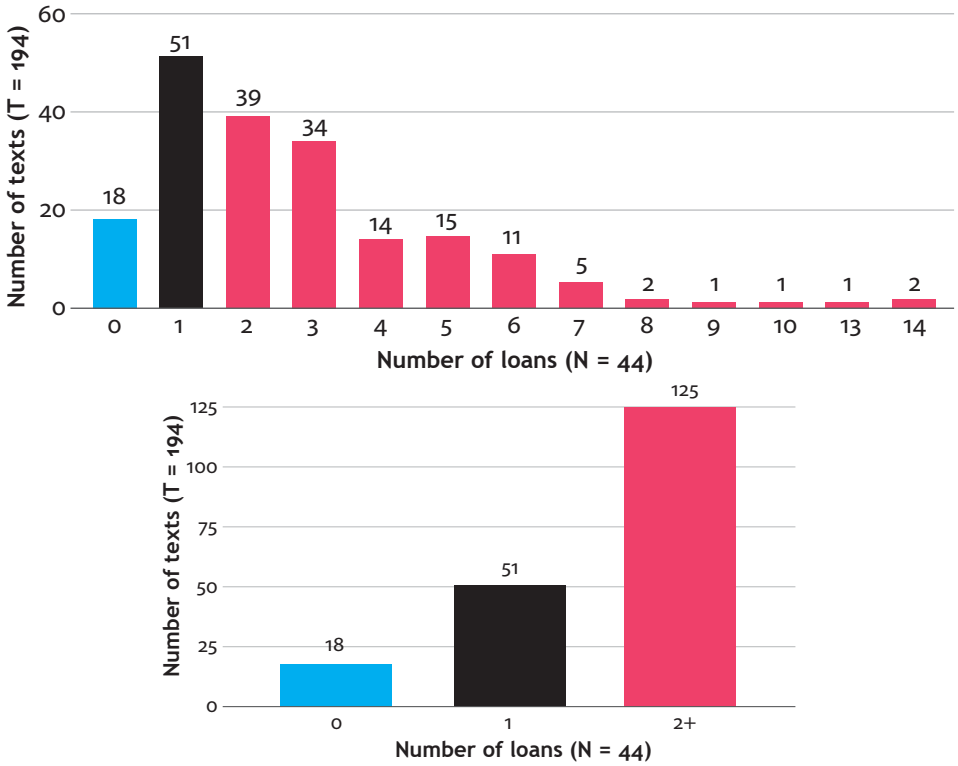


Figure 9.5: The number of loan types per text, including texts with no loans (blue) and one loan (black), which are omitted from our analysis.

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 black bar in Figure 9.5). 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.

9.4.2 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 9.6– 9.8). 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, beyond the topological effects. 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 9.6– 9.8 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 I) 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).

Looking at Figure 9.6, 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, except for 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"

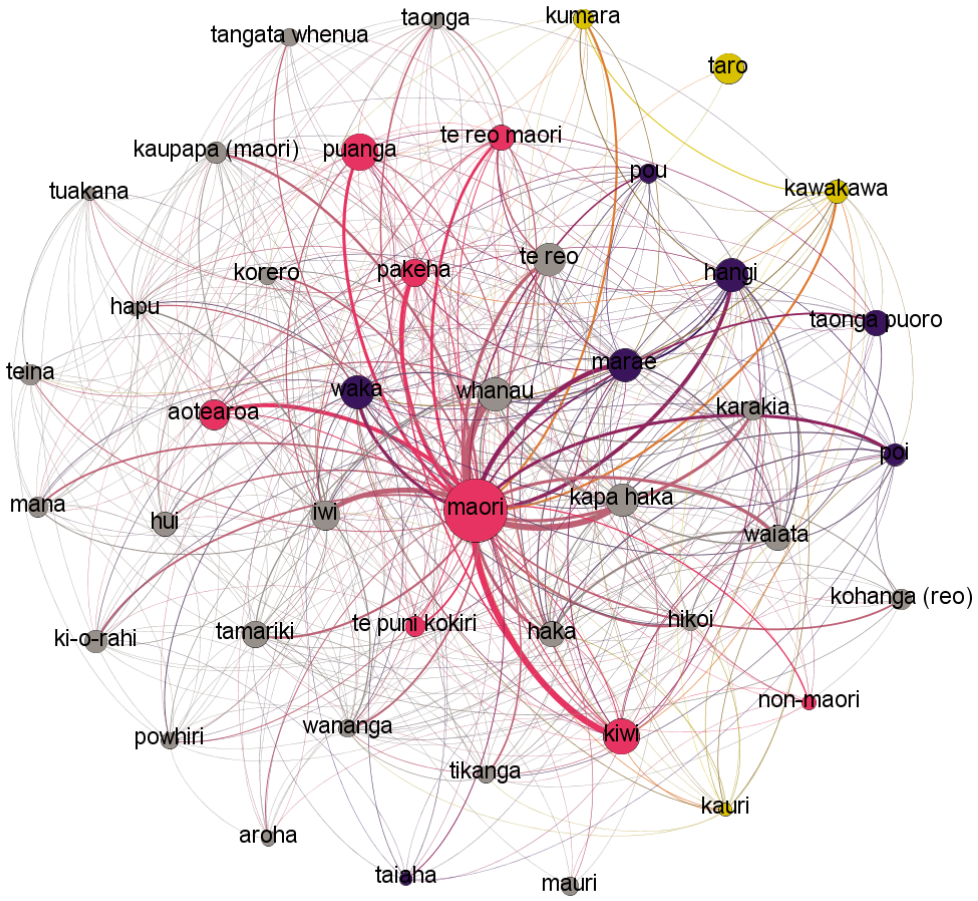


Figure 9.6: Standard network encoding semantic domain (social culture=grey, proper noun=pink, material culture=purple, flora and fauna=yellow).

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 9.7 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 9.8), namely that listed (and more familiar) loanwords are generally more central in the network, and unlisted (less familiar, possibly newer borrowings) are more peripheral.

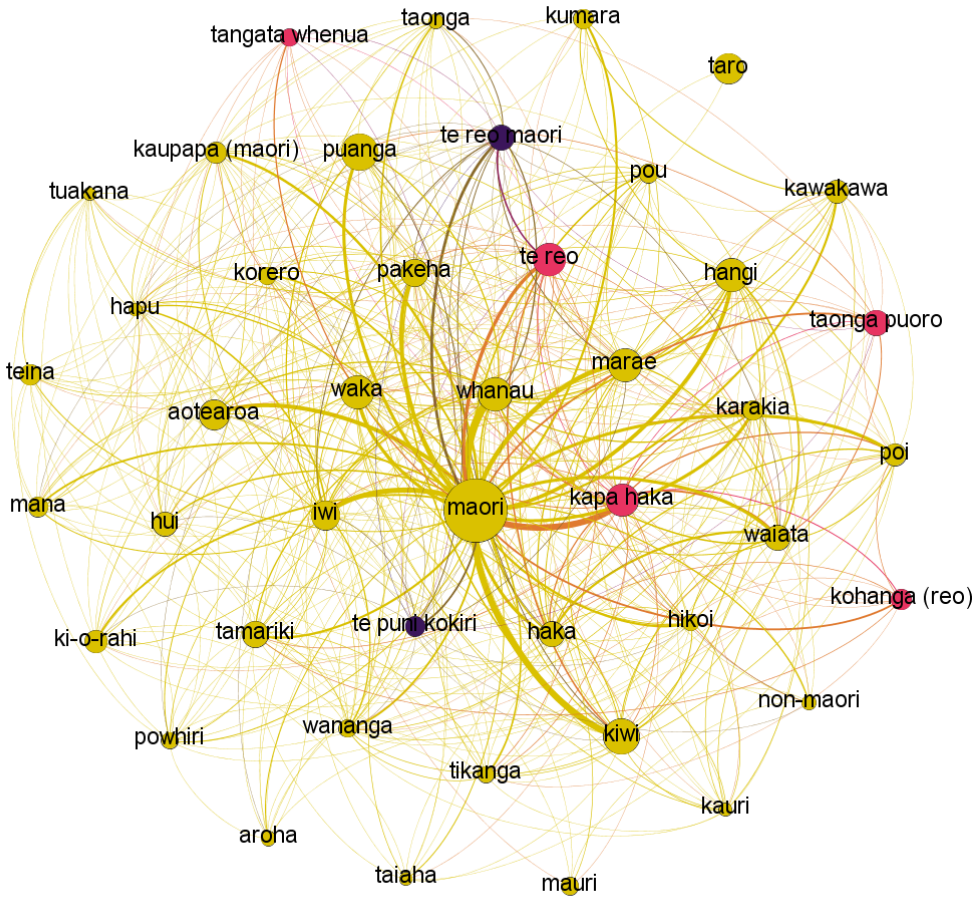


Figure 9.7: Standard network encoding loan size (one word=yellow, two words=pink, three words=purple).

Networks can also be explored from a statistical perspective, as given in Table 9.2. Apart from 'total edges', these metrics do not consider 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.

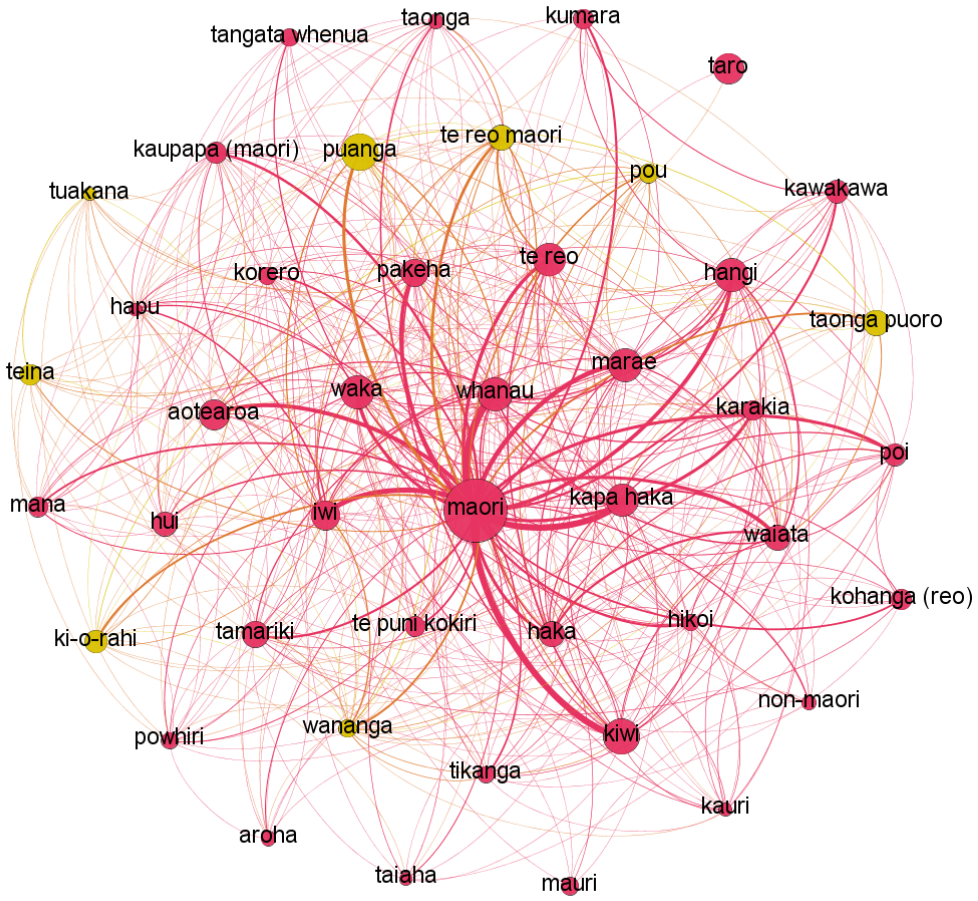


Figure 9.8: Standard network encoding listedness (listed=pink, unlisted=yellow).

Table 9.2: Network statistics for loans with at least five occurrences.

Metric	Value with *Māori* included (as per Figures 9.6– 9.8)	Value with *Māori* removed
Nodes	44	43
Distinct edges	483	440
Total edges	1,042	700
Average degree	22	20
Network density	51%	49%
Triadic closure	65%	64%

9.4.3 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 whether other loans are also present in those texts, and, if so, how many times each combination occurs. Figure 9.5 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'[5] (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 any number of 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., 2016). 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 piece of software that can be used to this end is the online tool *PAOHVis* (Valdivia et al., 2019; Aviz, 2022), 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 J), we have removed it from the following analysis, leaving 90 texts with two or more of the remaining 43 loans.

Figure 9.9 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 9.4, even infrequent loans appear to be widely spread – relative to their frequency – among texts containing multiple loans.

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

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

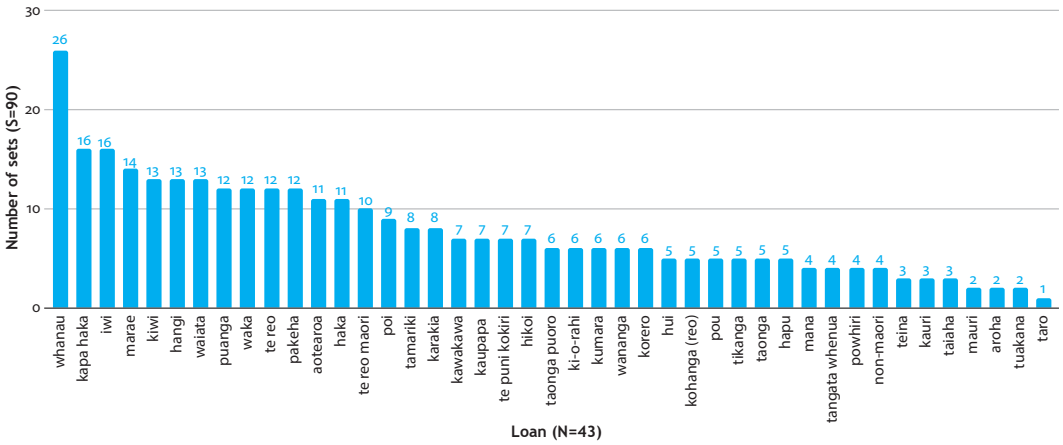


Figure 9.9: Loans by total number of sets (excluding the outlier *Māori*).

sets (n=49, 54.4%) contain at least one of the three most influential loans. Figures 9.10– 9.11 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 9.10, 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 9.9. 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 9.11 shows the same data, but with the nodes arranged by importance in their categories, enabling comparisons within and between the various groups.

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* are available online (Kiwi Words, 2021).

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

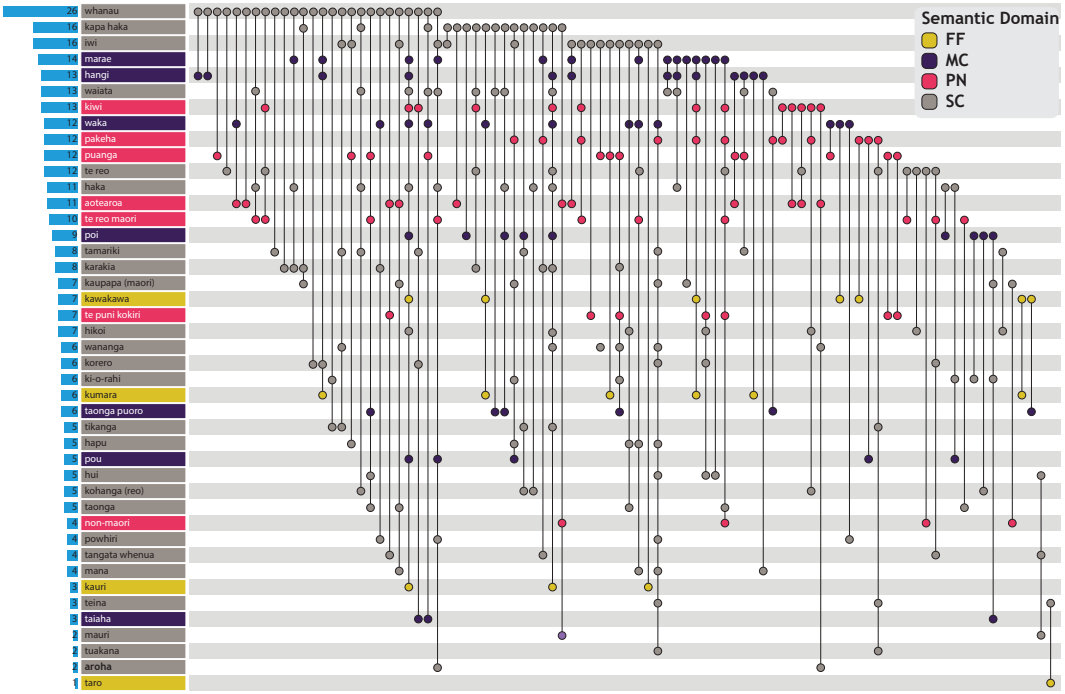


Figure 9.10: *PAOHVis* hypergraph showing all 90 sets coloured by semantic domain.

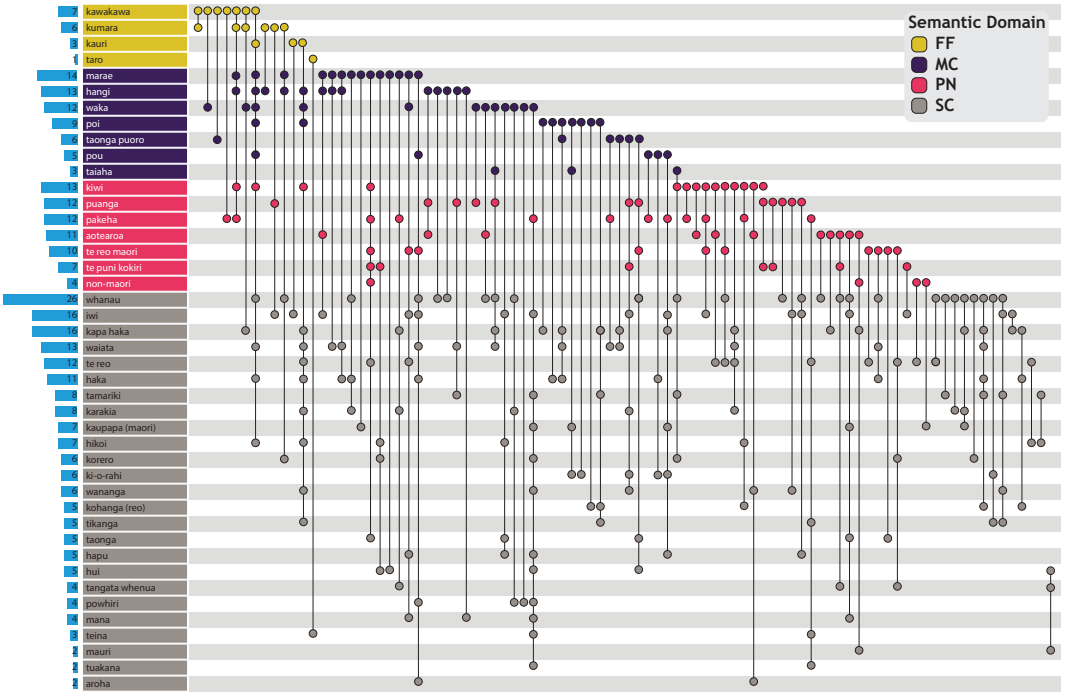


Figure 9.11: *PAOHVis* hypergraph, this time showing sets ordered by semantic domain.

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.[6]

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 eight, we again removed it from the 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 9.12 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 pink 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 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.

Social culture loans are the most prevalent category in Figure 9.12, 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 9.3.4. 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). Likewise,

[6]See Appendix G (pp. 349–352) for further details.

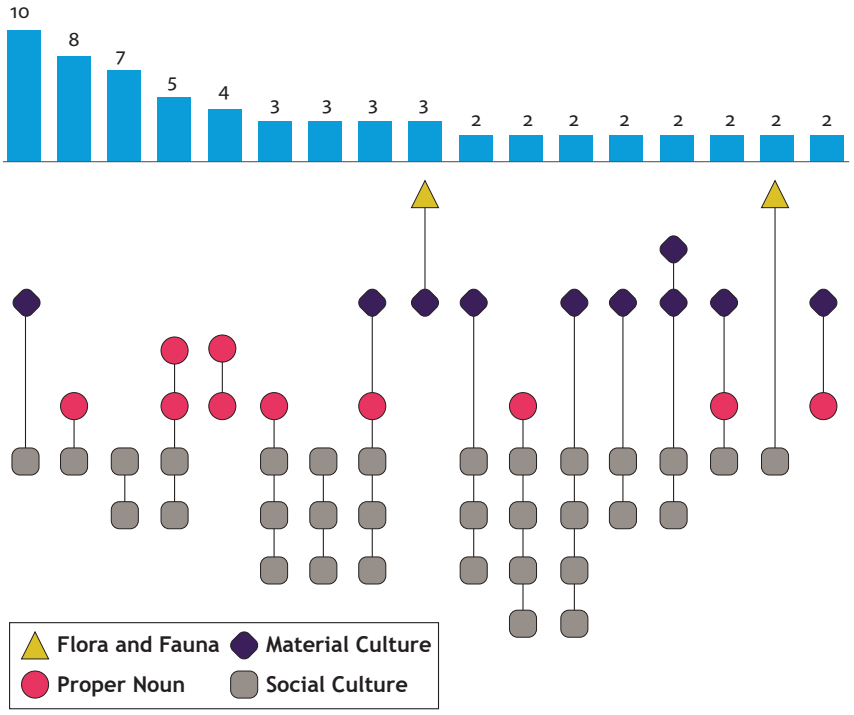


Figure 9.12: Loans aggregated by semantic domain (recurrent set configurations for 62 texts, 69% of the data).

there is usually a maximum of one material culture loan per text. 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 9.12 involve loans from just two of the four categories, and none has loans from all four categories. However, Figure 9.12 only shows set configurations that are shared by two or more texts, which applies to just over two-thirds of the data (the 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 9.13). 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 regarding the number of identical configurations they are part of. Flora and fauna terms are generally less dominant, the only

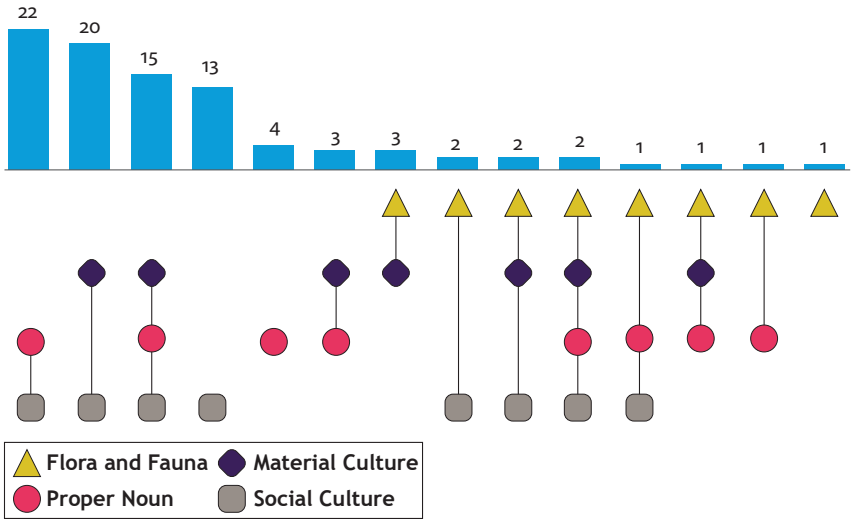


Figure 9.13: Loan sets collapsed by presence of semantic domain categories.

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 purple diamond in Figure 9.13.

The next set of visualisations (Figures 9.14– 9.15) show loans aggregated 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 9.14), 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 9.15, 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) always co-occur with at least one single-word loan in a text.

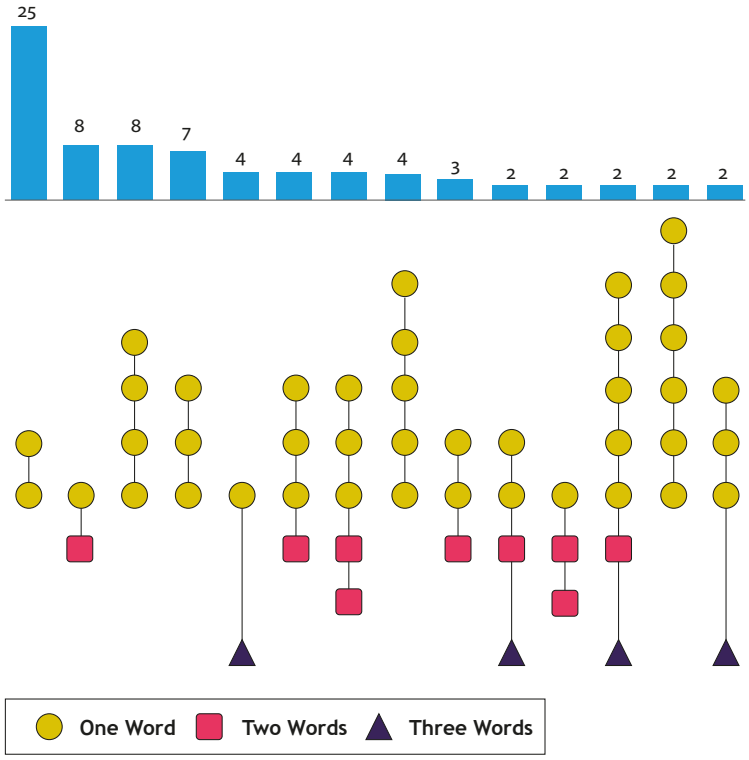


Figure 9.14: Loans aggregated by size (recurrent set configurations for 77 texts, 86% of the data).

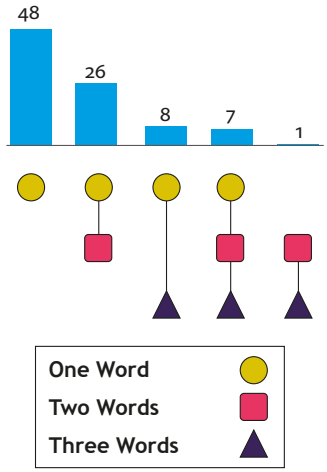


Figure 9.15: Loan sets collapsed by presence of size categories.

Figures 9.16– 9.17 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 9.16). 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

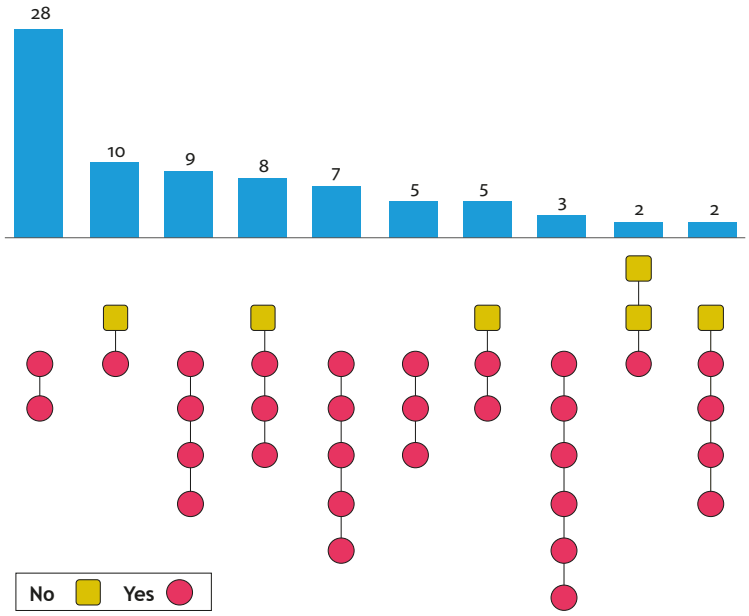


Figure 9.16: Loans aggregated by listedness (recurrent set configurations for 79 texts, 88% of the data).

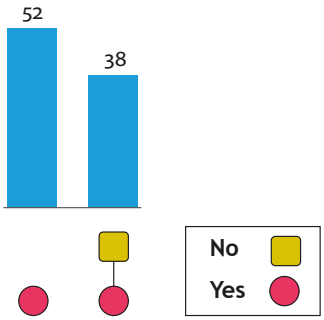


Figure 9.17: Loan sets collapsed by presence of listedness categories.

loans than listed ones. Although most sets consist solely of listed loans (58%), Figure 9.17 shows that there is still a large proportion of texts that contain a mixture of listed and unlisted items (42%).

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 band contains 10–12 items, based on the loans' raw frequency rankings (see Appendix H). Band 1 contains the most frequent loans and Band 4 the least frequent loans. According to Figure 9.18, none of the recurrent sets has 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

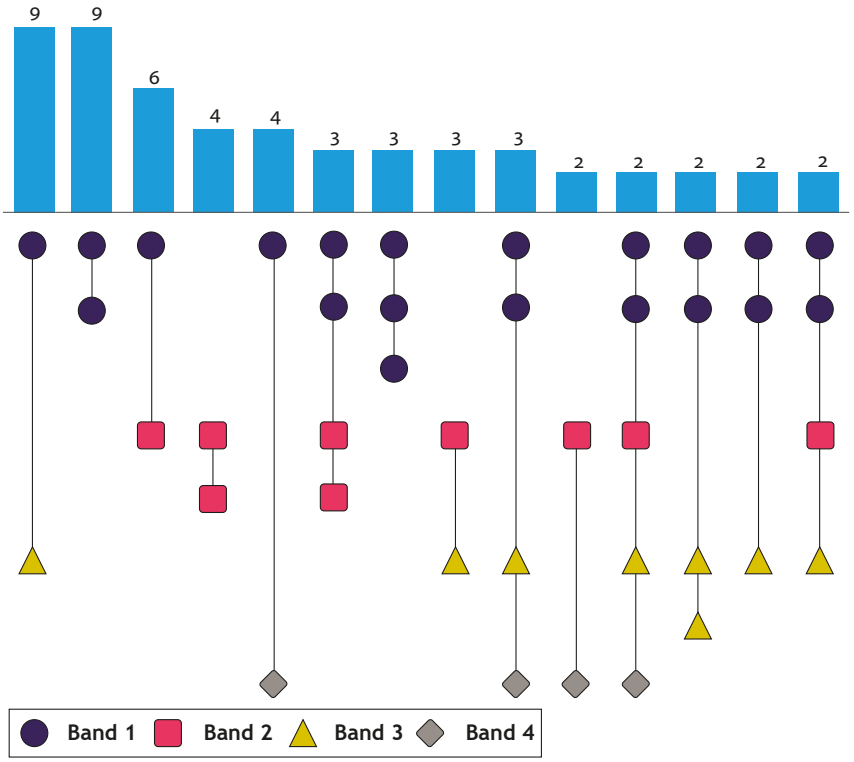


Figure 9.18: Loans aggregated by frequency bands (recurrent set configurations for 54 texts, 60% of the data).

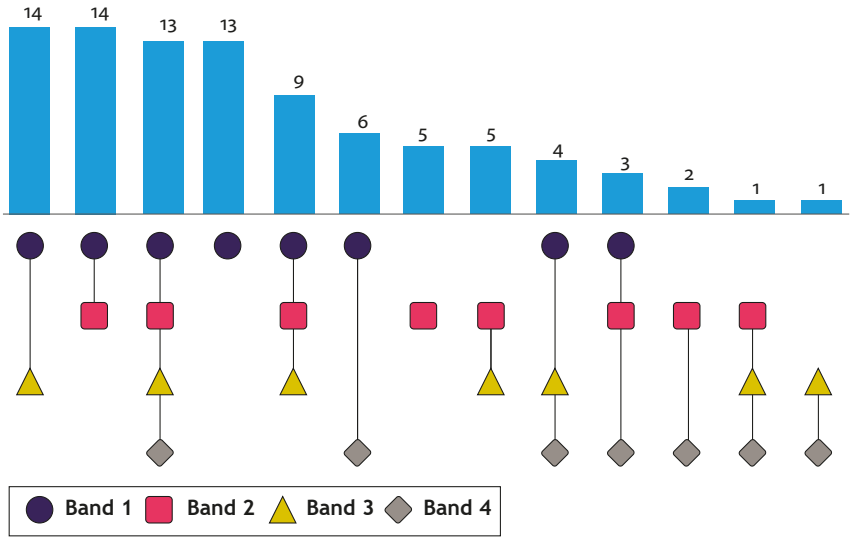


Figure 9.19: Loan sets collapsed by presence of frequency bands.

2, five from Band 3 and three from Band 4. Overall, each text appears to contain more loans from higher frequency bands.

Figure 9.19 considers only the presence or absence of each frequency band within a text. All texts except one contain loans from at least one 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, occurring in 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.

9.5 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 9.3, 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 most 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 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. Kurtböke and Potter (2000) limit their scope to collocates, looking to the immediate left or right of a given loan. Nevertheless, 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, both at micro- and macro-discourse levels.

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 recent years (e.g. Calude et al., 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 constitutes an example of the latter, as it is 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 it is true that 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, 2000), 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 limita-

tion is that the number of loans in each category is highly skewed, which means our results favour the more well-represented categories. This problem is partly due to our loanword selection criteria (see Section 9.3.2), 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.

9.6 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.

Funding

DT acknowledges funding from the University of Waikato Doctoral Scholarship. AC is grateful to the Royal Society of NZ Marsden Fund for their generous financial support.

Acknowledgements

We would like to thank the anonymous reviewers and the editor for their helpful comments and suggestions. We also thank members of the audience at LangSoc 2020 for their feedback on key ideas presented in this work.

9.7 Postscript

From an information visualisation perspective, this case study has provided an example of how categorical data can be fruitfully analysed when they occur with another data structure; it is not always appropriate to only consider categorical variables by themselves. We have shown how an existing technique intended for relational data (namely, PAOHVis) can be modified to enable the effective analysis of multiple categorical attributes in a stepwise fashion. It would also be beneficial to consider ways in which all of these attributes could be encoded at the same time, to explore multi-way interactions, without destroying the integrity of the underlying hypergraph representation.

From a corpus linguistics perspective, the chapter has shown how networks and hypergraphs can be used to uncover fresh insights into loanword use, especially when the loanwords are analysed in relation to other linguistic properties, such as semantic domain and listedness. The visualisations in this chapter, including the novel aggregated hypergraphs, revealed patterns that would not have been evident from employing traditional frequency-based methods for analysing loanword use. The networks highlighted pairwise relationships, while the hypergraphs preserved the loanword sets in their entirety.

Our findings suggest that Māori loanwords tend not to occur by themselves within an article; rather, authors who use one loanword often use several. The presence of both listed and unlisted loans in many articles indicates that New Zealand English is still in the process of importing new items from Māori, potentially signalling a third 'wave' of borrowing. Interestingly, after this research was carried out, a further 36 Māori words were added to the *Oxford English Dictionary*, underscoring the language's 'profound and lasting impact on English in New Zealand' (Salzar, 2023). Indeed, the representation and visibility of Māori words in the media and other domains both reflects and affects changes to social norms.

9.8 References

Aviz (2022). PAOHvis: Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization.

Bastian, M., Heymann, S., and 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*, volume 3, pages 361–362.

Berge, C. (1973). *Graphs and Hypergraphs*. North-Holland.

Calude, A., Stevenson, L., Whaanga, H., and Keegan, T. T. (2020a). The use of Māori words in National Science Challenge online discourse. *Journal of the Royal Society of New Zealand*, 50(4):491–508.

Calude, A. S., Miller, S., Harper, S., and Whaanga, H. (2019). Detecting language change: Māori loanwords in a diachronic topic-constrained corpus of New Zealand English newspapers. *Asia and Pacific Variation Journal*, 5(2):109–137.

Calude, A. S., Miller, S., and Pagel, M. (2020b). Modelling loanword success - a sociolinguistic quantitative study of Māori loanwords in New Zealand English. *Corpus Linguistics and Linguistic Theory*, 16(1):29–66.

Chesley, P. and Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, 48(6):1343–1374.

Davies, C. and 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 Facchinetti, R., Crystal, D., and Seidlhofer, B., editors, *From International to Local English–and Back Again*, pages 165–196. Peter Lang.

Denis, D. and D'Arcy, A. (2018). Settler colonial Englishes are distinct from postcolonial Englishes. *American Speech*, 93(1):1–31.

Deverson, T. and Kennedy, G. (2005). *The New Zealand Oxford Dictionary*. Oxford University Press.

Firth, J. R. (1957). *Papers in Linguistics*. Oxford University Press.

Fruchterman, T. M. and 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 University Press.

Görlach, M. (2002). *English in Europe*. 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. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11–15.

Hashimoto, D. (2019). Loanword phonology in New Zealand English: Exemplar activation and message predictability.

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 Moore, B., editor, *Who's Centric now? The Present State of Post-colonial Englishes*, pages 59–81. Oxford University Press.

Kiwi Words (2021). Loanword co-occurrence networks. `https://github.com/Waikato/kiwiwords/tree/master/loanword_networks`. [Accessed: 2023-04-23].

Kurtböke, P. and Potter, L. (2000). Co-occurrence tendencies of loanwords in corpora. *International Journal of Corpus Linguistics*, 5(1):83–100.

Laufer, B. and Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3):307–322.

Levendis, K. and Calude, A. S. (2019). Flagging loanwords and what they can tell us–a case study from New Zealand English. *Ampersand*, 6:100056.

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):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. and Daly, N. (2013). Kiwi, kapai, and kuia: Māori loanwords in New Zealand English children's picture books published between 1995 and 2005. In Carrington, B. and Pinsent, P., editors, *The Final Chapters: Concluding Papers of the Journal of Children's Literature Studies*, pages 44–56. Wizard's Tower Press.

Muysken, P. C. (2000). *Bilingual Speech: A Typology of Code-mixing*. Cambridge University Press.

Nobre, C., Meyer, M., Streit, M., and Lex, A. (2019). The state of the art in visualizing multivariate networks. *Computer Graphics Forum*, 38(3):807–832.

Onysko, A. and 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 University Press.

Qian, T., Ji, D., Zhang, M., Teng, C., and 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*, pages 1601–1611. Dublin City University and Association for Computational Linguistics.

Salzar, D. (2023). Words from the land of the long white cloud: New Zealand English additions to the OED. `https://www.oed.com/discover/words-from-the-land-of-the-long-white-cloud-new-zealand-/english-additions-to-the-oed`.

Sinclair, J. (1996). The search for units of meaning. *Textus*, 9:75–105.

Soriano-Morales, E., Ah-Pine, J., and Loudcher, S. (2016). Using a heterogeneous linguistic network for word sense induction and disambiguation. *Computación y Sistemas*, 20(3):315–325.

Stammers, J. R. and 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 on June 24, 2021.

Te Hiku Media (2019). Identify māori text. `https://github.com/TeHikuMedia/nga-kupu`.

Trye, D., Bravo-Marquez, F., Calude, A., and Keegan, T. T. (2020). Hybrid hashtags – #YouKnowYoureAKiwiWhen your tweet contains Māori and English. *Frontiers Special Issue on Computational Sociolinguistics*, 3.

Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N., and Fekete, J. D. (2019). Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE Transactions on Visualization and Computer*

Graphics, 27(1):1–13.

Weinreich, U. (1953). *Languages in Contact*. The Hague.

West, D. B. (1996). *Introduction to Graph Theory*, volume 2. Prentice hall.

Zenner, E., Speelman, D., and 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., and Geeraerts, D. (2013). What makes a catchphrase catchy? Possible determinants in the borrowability of English catchphrases in Dutch. In Zenner, E. and Kristiansen, G., editors, *New Perspectives on Lexical Borrowing*, pages 41–64. De Gruyter.

Zenner, E., Speelman, D., and 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, G. K. (1935). *The Psychobiology of Language*. Houghton-Mifflin.

Part IV

Thesis Conclusion

Chapter 10

Conclusion

This thesis has developed and adapted visualisation techniques for representing multivariate categorical data, and has applied these to corpus linguistics to demonstrate their effectiveness. In this concluding chapter, we discuss the overall contribution of the published papers and manuscripts, and identify key challenges, limitations and future work. We begin by restating the two overarching research questions posited in Chapter 1:

1. What generalisable information visualisation techniques can be developed or adapted to enable the effective analysis of datasets involving multiple categorical variables?

2. How can applying these techniques to a particular domain increase understanding of that domain?

10.1 Research Question One

The goal of the first research question was to extend the state of the art in visualising datasets involving multiple categorical variables. After providing necessary background information in Part I, this objective was addressed in Part II, beginning with a structured review of categorical data visualisation in Chapter 3. Six different families of categorical visualisation techniques were identified, each with their own strengths and weaknesses. Ultimately, this review confirmed the need for more scalable and interactive solutions for dealing with categorical data.

To address this identified gap, Chapter 4 presented extensions to the Heatmap Matrix and Chapter 5 introduced a novel aggregation-based technique called MultiCat. Drawing from the taxonomy from Chapter 3, both these techniques are 'table' representations belonging to different sub-categories: the Heatmap

Matrix Explorer is an example of a 'pairwise matrix', while MultiCat is a 'tabular' technique. Due to the bivariate nature of the Heatmap Matrix Explorer and the multivariate nature of MultiCat, these techniques can provide distinct but complementary insights into complex patterns and trends in a dataset. They additionally offer advanced interaction, and can handle a greater number of categories than other techniques identified in Chapter 3, making them useful tools for visualising multidimensional categorical data. We summarise key details about each of these techniques in turn, before reviewing how they were applied to the domain of corpus linguistics, in accordance with our second research question.

10.1.1 Heatmap Matrix Explorer

Compared with its predecessor (Rocha and da Silva, 2018), the Heatmap Matrix Explorer improves aesthetic and cognitive aspects of the main display; this can be seen in Figure 10.1. The Heatmap Matrix Explorer also incorporates controls for filtering and manually reordering variables and categories, as shown in Figure 10.2. Additionally, it provides a link to the underlying data through its coordinated table view, allowing the analyst to examine selected records in full.

Another key change is that the display integrates several cell-level metrics beyond raw frequency, including Pearson residuals and conditional (row/column) relationships. Two of these metrics can be encoded at the same time using a bivariate colour scheme. Furthermore, where criteria for the Chi-square test are met, associations among pairs of variables can be directly visualised in the heatmap. This feature highlights the potential for integrating statistical tests into categorical visualisation techniques more widely.

Currently, the Heatmap Matrix Explorer accommodates roughly 40 categories at the same time, regardless of how these are distributed among variables. This limit could potentially be extended to hundreds of categories[1] by: (i) enabling different zoom levels, and using distortion techniques to magnify selected regions; (ii) incorporating vertical and horizontal scroll bars in the matrix view; and (iii) adding functionality to collapse and expand the surrounding components.

[1]This is comparable to a standard heatmap encoding two variables, where the size of each cell can be reduced to a single pixel (Munzner, 2014, p. 159).

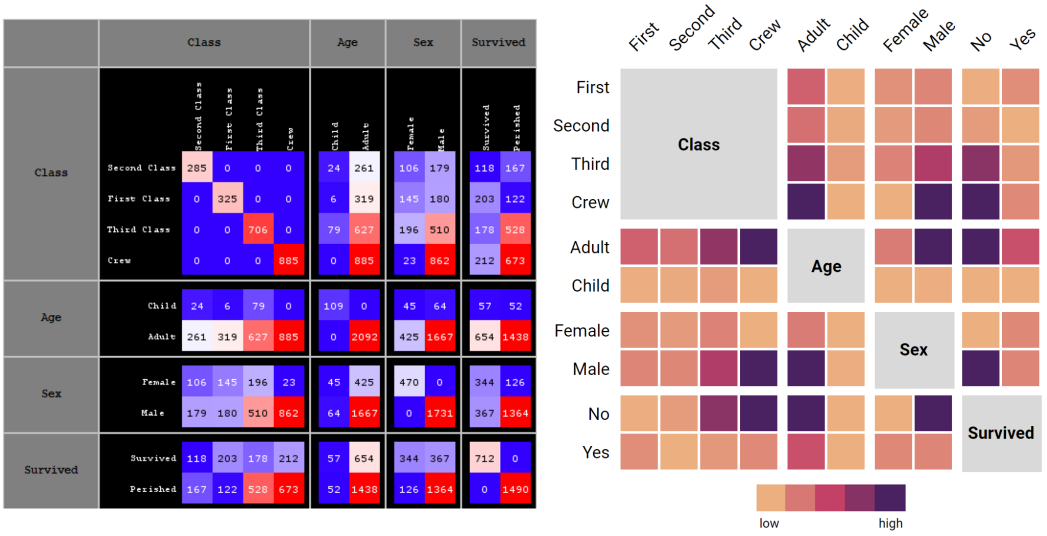


Figure 10.1: Side-by-side comparison of the original Heatmap Matrix (left) and the updated Heatmap Matrix Explorer (right), featuring the Titanic dataset (Dawson, 1995).

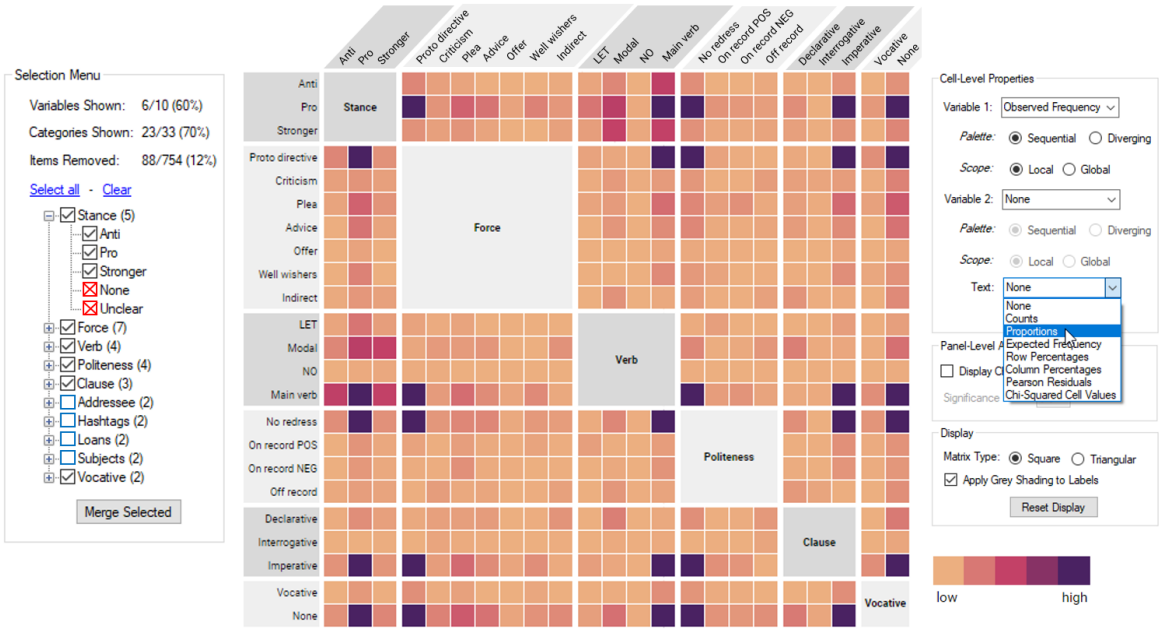


Figure 10.2: Menu controls for the Heatmap Matrix Explorer, reproduced from Figure 4.6. The coordinated table view is omitted because no cells have been selected. The dataset comprises directives used in tweets containing the hashtag *#covid19nz* (Burnette and Calude, 2022).

10.1.2 MultiCat

MultiCat, pictured in Figure 10.3, adopts a different approach to visualising multivariate categorical data, whereby rows represent aggregated combinations of categories and columns correspond to different variables.[2] This layout draws on users' familiarity with spreadsheets, avoiding line crossings and spatial regions—like tiles in a Mosaic Plot—whose length and height may vary considerably, making comparisons difficult. Interaction is also used to highlight the proportion of each category with respect to selected subgroups.

The MultiCat technique was evaluated via a small-scale user study, which yielded positive feedback. Participants successfully performed a wide range of tasks, from comparing category frequencies to identifying prominent multivariate relationships and computing *a priori* probabilities for specific subgroups of categories.

MultiCat handles datasets with up to 20 variables, each of which may contain any number of categories. However, variables with fewer than seven categories can be more readily distinguished, as additional instances are simply shown in grey. Unlike the Heatmap Matrix, MultiCat treats ordinal variables differently from nominal ones, by using greyscale colours and sorting categories according to their natural ordering. These design decisions for ordinal data were informed by visualisation theory, but their utility within MultiCat has not yet been formally evaluated.

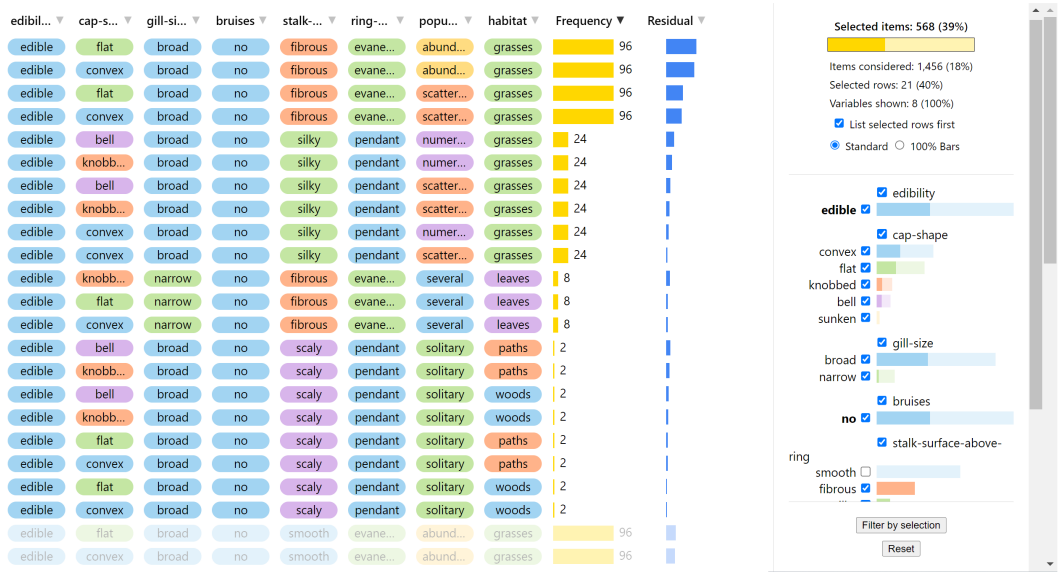


Figure 10.3: The MultiCat technique showing a filtered subset of the Mushroom dataset (Schlimmer, 1987).

[2]See `https://dgt12.github.io/multicat/` for a live demo.

10.2 Research Question Two

The goal of the second research question was to use the newly developed visualisation techniques to increase understanding of a specific application domain. This objective was fulfilled in Part III by applying the above techniques—and others—to language data, in order to gain deeper insight into how te reo Māori is used on Twitter (Case Study I, Chapter 8) and in NZE newpsaper articles (Case Study II, Chapter 9). In order to increase language resources for Māori, a new dataset was created for the first case study. This was accomplished using NLP techniques, as discussed in Chapter 6, which also led to the development of a second corpus comprising mixed Māori-English tweets (Chapter 7).

10.2.1 Case Study I: Māori Possessives

In the first case study, MultiCat was instrumental in examining typical semantic characteristics of possessive phrases in Māori-language tweets. It was easy to isolate and compare sets of possessive phrases that did and did not conform with expected usage. Furthermore, MultiCat proved valuable for identifying mistakes and inconsistencies in our annotations, which were then resolved. Interestingly, our data did not align with the recommended parameters for using MultiCat that were specified in Chapter 5, namely that it is best suited for representing many variables with low cardinalities. Instead, our dataset comprised relatively few variables with high cardinalities; this resulted in many grey categories but did not hinder our analysis. Additionally, the Heatmap Matrix Explorer was employed to investigate the characteristics of tweeters in the corpus, providing an overview of six sociolinguistic variables organised into 15 categories.

It is important to note that the static figures presented in Chapter 8 do not capture the interactive nature of these tools, such as the ability to hover over variable boxes in the Heatmap Matrix Explorer to display the marginal frequency of each category, or to rapidly encode different cell-level metrics within the heatmap itself. Unfortunately, the independence requirement for the Chi-square test poses a challenge in corpus linguistic studies, where multiple texts are often written by the same individuals. Indeed, this requirement was not met in our study, which is why the Chi-square test was not reported on within the paper.

10.2.2 Case Study II: Māori Loanword Co-occurrence

The second case study examined the co-occurrence of Māori loanwords in NZE newspaper articles in the existing Matariki Corpus (Calude et al., 2019). This involved visualising relational data with categorical attributes, rather than categorical data by itself. It was therefore necessary to use tools other than MultiCat and the Heatmap Matrix Explorer. In particular, network and hypergraph visualisations were employed to highlight different aspects of the data: the networks emphasised the highly interconnected nature of the loanwords, while the hypergraphs showed the exact nodes present within each text. Moreover, the hypergraphs were aggregated in new ways by leveraging each categorical attribute in succession, aligning with the first research question.

The aggregated hypergraphs helped to uncover findings that were not evident from the networks alone, such as the presence of one or more single-word borrowings in every text, and the frequent mixing of listed and unlisted loanwords. These hypergraphs also promoted further exploration of the data, including the binning of loanwords by frequency, so that this could be treated as an additional categorical variable for aggregation. This in turn led to the discovery that rare loanwords never occur by themselves, enriching our understanding of their use within the corpus.

10.2.3 Applications in Linguistics and Beyond

While these case studies are limited to specific features of two languages, the visualisation techniques employed are broadly applicable to other linguistic datasets involving multivariate categorical data, or relational data with categorical attributes. For instance, MultiCat could facilitate analysis of cross-linguistic patterns based on a carefully chosen subset of structural features from the Grambank database (Skirgård et al., 2023). In this case, each data item would represent a distinct language variety and each variable would correspond to a different phonological, grammatical or lexical feature. It would be prudent to include only languages that have been coded for all (or most) of the features of interest, so as to minimise 'artificial' combinations arising from missing values.

Zooming out further, although this thesis has focused primarily on linguistic applications, the proposed visualisation techniques offer practical value for analysts, in any domain, who frequently work with categorical data. Both MultiCat and the Heatmap Matrix Explorer are useful for gaining an overview of a dataset, detecting anomalies, identifying trends and testing hypotheses. They

excel at frequency-based tasks, explicitly supporting the visualisation of joint and conditional relationships. The aggregated hypergraphs, in comparison, simplify the analysis of large groups of entities comprising one or more categorical attributes. What these techniques have in common, besides their focus on categorical data, is the ability to uncover patterns that might otherwise go unnoticed.

10.3 Summary of Contributions

The main contributions of this thesis can be summarised as follows:

1. A review of existing categorical visualisation techniques:
    - An online database for exploring 120 relevant papers: `https://cat-vis.github.io`.
    - A taxonomy organising categorical visualisation techniques into six distinct families and a number of sub-families.
    - An overview of nine types of analysis tasks associated with visualising categorical data.
2. Generalisable visualisation techniques that aid the effective analysis of datasets involving multiple categorical variables:
    - Extensions to the pairwise Heatmap Matrix, realised in an empirical prototype called the Heatmap Matrix Explorer.
    - The design, implementation and evaluation of MultiCat, which facilitates the analysis of higher-order categorical relationships.
    - Extensions for aggregating PAOHVis hypergraphs (Valdivia et al., 2019) based on one or more categorical attributes.
3. Language resources for te reo Māori and the mixing of Māori and English:
    - The first publicly available corpus of Māori-language tweets, called the RMT Corpus, which can be used for linguistic analysis and the development of NLP resources (Chapter 6).
    - A hybrid architecture for labelling mixed Māori-English text, which combines linguistic rules with machine learning (Chapter 7).
    - The first labelled corpus of mixed Māori-English tweets, called the MET Corpus, which affords opportunities for better understanding language-contact-induced variation in Aotearoa New Zealand (Chapter 7).

- A systematic approach to classifying possessive phrases in Māori based on semantic criteria (Chapter 8).

4. Linguistic findings about possession in Māori-language tweets and loanword use in New Zealand English newspaper articles:

   - Empirical evidence confirming the dominance of the O category in Māori-language tweets, and a tendency for tweeters to use an *o* marker in situations where grammars would instead document *a* (Chapter 8).
   - Co-occurrence data that suggest Māori loanwords tend not to occur in texts by themselves, and that new items are still being borrowed from Māori (Chapter 9).

10.4 Challenges, Limitations and Future Work

The research presented in this thesis helps highlight current challenges and, further, suggests a number of avenues for future work. We summarise five of these below:

1. **Heterogeneous Datasets:** The techniques presented in Part II relate to purely categorical data, but many datasets contain a mixture of categorical and continuous variables. Although continuous variables can be binned, and subsequently represented using categorical techniques, this invariably results in loss of information. More flexible solutions are needed that can handle mixed datasets with a large number of categorical variables.

2. **User Studies:** The field of categorical data visualisation would benefit from more user studies, including comparative evaluations of different techniques to better understand their advantages and disadvantages with respect to specific tasks and datasets. The analysis tasks and technique taxonomy outlined in Chapter 3 would serve as a framework for such studies.

3. **Interactive Visualisation Tools**: There are currently few well-maintained, interactive tools available for visualising multivariate categorical data. To address this gap, there is a need for additional open-source implementations that do not require coding skills from the end user. This would help to make categorical visualisation techniques more accessible to non-computer scientists.

4. **NLP Tools for Māori and NZE:** New and improved NLP tools that minimise bias against Māori and NZE are crucial, particularly in light of recent developments in large language models. The corpora discussed in Chapters 5 and 6 provide valuable training data for the refinement of such tools, particularly with regard to informal language use.

5. **Quantitative Analyses of Māori Grammar:** Building on Chapter 8, further empirical analyses of other aspects of Māori grammar would be beneficial for developing usage-based language resources that support both teachers and learners of the Indigenous language of Aotearoa. Such studies could be undertaken using the RMT Corpus, and supplemented with datasets from other genres.

10.4.1 Limitations

Regarding limitations of the visualisation techniques presented in this thesis, the Heatmap Matrix Explorer has yet to undergo formal evaluation or be implemented as a fully-functional web-based tool. Our empirical prototype does not incorporate automatic methods for arranging categories and variables in a data-driven manner; it merely displays them in the order they appear in the raw data. It would be advantageous to process ordinal variables in a similar manner to MultiCat, so that the intrinsic order of these categories can be used by default. This would also facilitate the application of specialised tests for pairs of ordinal variables, in lieu of the Chi-square test, which was designed for nominal data.

Likewise, MultiCat has some limitations as currently implemented, which could be addressed in future work. Although the online tool showcases the core functionality of the technique, it lacks the more advanced features discussed in Section 5.11. In particular, we believe the ability to disaggregate combinations and view the corresponding records, like in Taggle (Furmanova et al., 2020), would better meet user expectations for a spreadsheet-like tool. This could be effectively accompanied by a search feature to browse specific records of interest.

While MultiCat and the Heatmap Matrix Explorer allow for unknown values, neither currently supports analysis tasks related to missing data, such as summarising the distribution of missing values across all variables, or removing incomplete data in bulk. However, adding such functionality could be achieved with minimal changes to the existing menu controls. Furthermore, it can be useful to explicitly visualise all combinations of categories that do *not* occur

in a dataset, as people tend to focus on immediately observable phenomena. In the Heatmap Matrix Explorer, non-observed combinations are indicated by zero-cells, which can be highlighted in a different colour. MultiCat, however, does not currently display such combinations. Introducing a checkbox to toggle the display of non-observed combinations would be a straightforward feature to integrate.

The aggregated hypergraphs from Chapter 9 have not been implemented in a web-based tool, highlighting another opportunity for future work. We envisage an extension to PAOHVis (Valdivia et al., 2019) that would allow users to select one or more categorical attributes for aggregation (e.g., via drop-down menus or drag-and-drop panes for different visual channels). The data would be aggregated using the count-based approach first, followed by the binary approach, with the option to dynamically switch back and forth between aggregation levels, including restoring the original hypergraph. Figure 10.4 provides a conceptual illustration, adapted from Appendix G. Animated transitions (Heer and Robertson, 2007) could be used to enhance the user's understanding of changes between aggregation levels.

It is important to note that our suggested methods for aggregation work best when based on a single categorical attribute. While multiple visual channels such as shape, texture and colour can be employed to aggregate a hypergraph according to multiple categorical variables, this typically reduces the number of sets with identical properties, limiting the extent to which the visualisation can be simplified. When the hypergraph is not aggregated, however, mapping multiple categorical attributes in this way may be feasible.

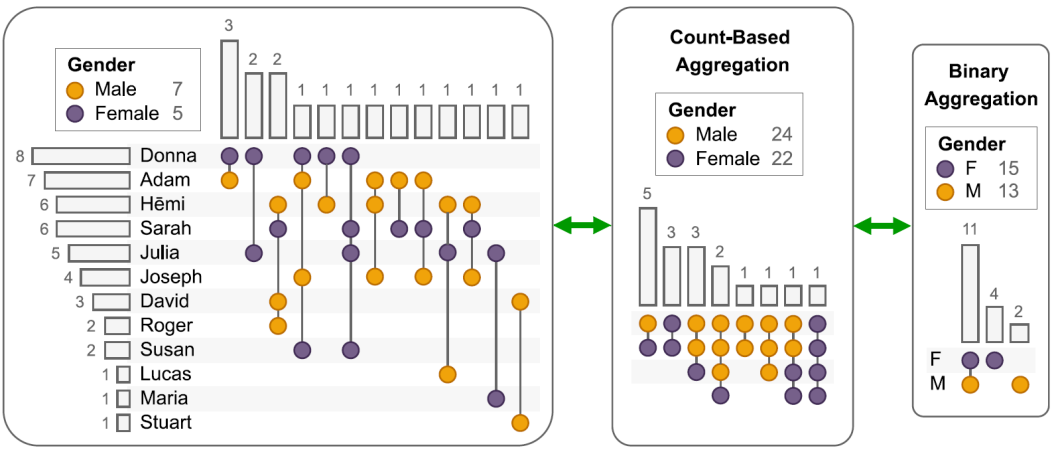


Figure 10.4: Proposed aggregation levels for PAOHVis hypergraphs, adapted from Appendix G. The figure displays fictitious co-authorship data for 12 individuals, with gender being the categorical attribute driving the aggregation.

10.4.2 Closing Remarks

Although not an explicit requirement at the inception of this research, an important recurring feature has been the combined use of aggregation and interaction to aid users in exploring complex data. Interaction has repeatedly proven to be a powerful mechanism, particularly in the context of multivariate data, by highlighting how the different variables and views relate to or influence one another. The techniques presented in this thesis actively embrace the idea that no single *static* representation can achieve everything a group of users could possibly need.

To further build on this foundation, MultiCat could be enhanced to support multi-level hierarchical categorical data using a *top-down* approach (Elmqvist and Fekete, 2009), similar to the one adopted by Vosough et al. (2018). For instance, clicking on a plus icon next to a column header in MultiCat could allow users to drill down to the next level of the hierarchy for that particular variable (e.g., ten-year age bands for Titanic passengers, rather than just 'adult' and 'child' categories), with all combinations being updated accordingly. Conversely, clicking on a minus sign for the same variable would 'roll up' to the previous state, facilitating rapid traversal of the data's tree structure in both directions and across multiple levels.

This is not merely a process of aggregating or disaggregating the data once and concluding there; rather, it provides the flexibility for users to continually navigate back and forth according to their needs, in an entirely consistent manner. However, while enriching exploration, this approach could lead to cognitive overload in MultiCat, due to the explosion of categories and combinations that may occur as one drills further down. Indeed, this issue is similar to the challenges encountered when aggregating hypergraphs by multiple attributes, as described above.

There is scope to evolve the visualisation techniques presented in this thesis into more robust visual analytics tools. Visual analytics combines the strengths of visualisation, interaction and automated analysis methods (Keim et al., 2008), yet this thesis has primarily addressed only the first two components. Automated analyses are helpful as they facilitate identification of features or abstractions that simplify the interpretation of raw data (Tominski and Schumann, 2020). For example, the tools discussed could be enhanced by integrating automatic (but user-verified) extraction of the most relevant variables at the beginning of the analysis. Additionally, during the visualisation process, it may be beneficial to provide options to calculate, rank and visualise categories, variables, combinations or data items based on various

similarity measures, such as Cosine distance, Mutual Information (MI) or the Jaccard index. This would help to compensate for the fact that CatVis techniques are geared towards frequency-based tasks rather than similarity tasks (Johansson Fernstad and Johansson, 2011).

Finally, throughout all stages of the analysis, automatic suggestions could effectively guide the user in their exploration. Drawing inspiration from Wongsuphasawat et al. (2015) and Gu et al. (2024), non-obtrusive recommendations could be provided to the user, together with the option to confirm, modify, ignore or disable these. This approach seeks to optimise the synergy of human-in-the-loop and computational power, where machines augment—but do not supplant—people's knowledge, creativity and sense-making processes.

10.5 References

Burnette, J. and Calude, A. S. (2022). Wake up New Zealand! Directives, politeness and stance in Twitter #covid19nz posts. *J. Pragmat.*, 196:6–23.

Calude, A. S., Miller, S., Harper, S., and Whaanga, H. (2019). Detecting language change: Māori loanwords in a diachronic topic-constrained corpus of New Zealand English newspapers. *Asia and Pacific Variation Journal*, 5(2):109–137.

Dawson, R. J. M. (1995). The "unusual episode" data revisited. *Journal of Statistics Education*, 3(3).

Elmqvist, N. and Fekete, J.-D. (2009). Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE transactions on visualization and computer graphics*, 16(3):439–454.

Furmanova, K., Gratzl, S., Stitz, H., Zichner, T., Jaresova, M., Lex, A., and Streit, M. (2020). Taggle: Combining overview and details in tabular data visualizations. *Information Visualization*, 19(2):114–136.

Gu, K., Grunde-McLaughlin, M., McNutt, A., Heer, J., and Althoff, T. (2024). How do data analysts respond to AI assistance? A Wizard-of-Oz study. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–22.

Heer, J. and Robertson, G. (2007). Animated transitions in statistical data graphics. *IEEE transactions on visualization and computer graphics*, 13(6):1240–1247.

Johansson Fernstad, S. and Johansson, J. (2011). A task based performance evaluation of visualization approaches for categorical data analysis. In *2011 15th International Conference on Information Visualisation*, pages 80–89. IEEE.

Keim, D., Andrienko, G., Fekete, J.-D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). *Visual analytics: Definition, process, and challenges*. Springer.

Munzner, T. (2014). *Visualization analysis and design*. CRC press.

Rocha, M. and da Silva, C. G. (2022). Heatmap Matrix: Using reordering, discretization and filtering resources to assist multidimensional data analysis.

Rocha, M. M. N. and da Silva, C. G. (2018). Heatmap Matrix: a multidimensional data visualization technique. In *Proceedings of the 31st Conference*

on Graphics, Patterns and Images (SIBGRAPI).

Schlimmer, J. (1987). Mushroom. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5959T.

Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Chira, A., Maurits, L., Dinnage, R., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., Robbeets, M., Abbas, N. K., Auer, D., Bakker, N. A., Barbos, G., Borges, R. D., Danielsen, S., Dorenbusch, L., Dorn, E., Elliott, J., Falcone, G., Fischer, J., Ghanggo Ate, Y., Gibson, H., Göbel, H.-P., Goodall, J. A., Gruner, V., Harvey, A., Hayes, R., Heer, L., Herrera Miranda, R. E., Hübler, N., Huntington-Rainey, B., Ivani, J. K., Johns, M., Just, E., Kashima, E., Kipf, C., Klingenberg, J. V., König, N., Koti, A., Kowalik, R. G. A., Krasnoukhova, O., Lindvall, N. L., Lorenzen, M., Lutzenberger, H., Martins, T. R., Mata German, C., van der Meer, S., Montoya Samamé, J., Müller, M., Muradoğlu, S., Neely, K., Nickel, J., Norvik, M., Oluoch, C. A., Peacock, J., Pearey, I. O., Peck, N., Petit, S., Pieper, S., Poblete, M., Prestipino, D., Raabe, L., Raja, A., Reimringer, J., Rey, S. C., Rizaew, J., Ruppert, E., Salmon, K. K., Sammet, J., Schembri, R., Schlabbach, L., Schmidt, F. W., Skilton, A., Smith, W. D., de Sousa, H., Sverredal, K., Valle, D., Vera, J., Voß, J., Witte, T., Wu, H., Yam, S., Ye , J., Yong, M., Yuditha, T., Zariquiey, R., Forkel, R., Evans, N., Levinson, S. C., Haspelmath, M., Greenhill, S. J., Atkinson, Q. D., and Gray, R. D. (2023). Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances*, 9.

Tominski, C. and Schumann, H. (2020). *Interactive visual data analysis*. AK Peters/CRC Press.

Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N., and Fekete, J. D. (2019). Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(1):1–13.

Vosough, Z., Hogräfer, M., Royer, L. A., Groh, R., and Schulz, H.-J. (2018). Parallel hierarchies: A visualization for cross-tabulating hierarchical categories. *Computers & Graphics*, 76:1–17.

Wongsuphasawat, K., Moritz, D., Anand, A., Mackinlay, J., Howe, B., and Heer, J. (2015). Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics*, 22(1):649–658.

Appendix A

Co-Authorship Forms

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work**. Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 4: Extending the Heatmap Matrix: Pairwise Analysis of Multivariate Categorical Data
2023 27th International Conference on Information Visualisation (IV)
Pp. 29-36. Tampere, Finland: IEEE

Nature of contribution by PhD candidate	Conceptualised the research; created the prototype; wrote the draft.
Extent of contribution by PhD candidate (%)	80

CO-AUTHORS

Name	Nature of Contribution
Mark Apperley	Provided guidance and critical feedback
David Bainbridge	Provided guidance and critical feedback

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

Name	Signature	Date
Mark Apperley		24/11/23
David Bainbridge		24/11/23

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.
Chapter 5: MultiCat: A Visualisation Technique for Multidimensional Categorical Data (Unpublished manuscript)

Nature of contribution by PhD candidate	Conceptualised the research; created the prototype; conducted the user study; wrote the draft
Extent of contribution by PhD candidate (%)	80

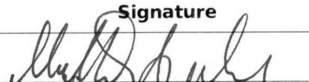
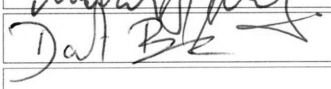
CO-AUTHORS

Name	Nature of Contribution
Mark Apperley	Provided guidance and critical feedback
David Bainbridge	Provided guidance and critical feedback

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

Name	Signature	Date
Mark Apperley		01 Feb 2024
David Bainbridge		1/2/24

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Mataurauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.	Chapter 6: Harnessing Indigenous Tweets: The Reo Māori Twitter Corpus Language Resources and Evaluation, 56(4), 1229-1268.
Nature of contribution by PhD candidate	Curated, tagged and analysed the data; wrote the first draft
Extent of contribution by PhD candidate (%)	85

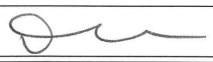
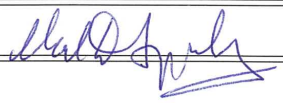
CO-AUTHORS

Name	Nature of Contribution
Te Taka Keegan	Provided guidance and critical feedback
Paora Mato (deceased)	Provided guidance and critical feedback
Mark Apperley	Provided guidance and critical feedback

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

Name	Signature	Date
Te Taka Keegan		23/11/23
Mark Apperley		24/11/23

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Mataurange Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.
Chapter 7: A Hybrid Architecture for Labelling Bilingual Māori-English Tweets
Findings of the Association for Computational Linguistics: NAACL-IJCNLP 2022
pp. 119-130

Nature of contribution by PhD candidate	Collected, annotated and analysed the data; jointly developed the architecture; wrote most of the first draft.
Extent of contribution by PhD candidate (%)	70

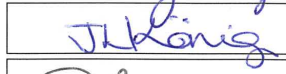
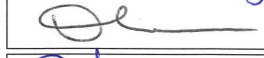
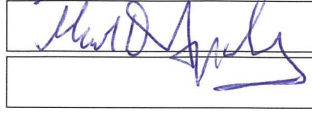
CO-AUTHORS

Name	Nature of Contribution
Vithya Yogarajan	Developed the architecture; conducted and wrote about the ML experiments
Jemma König	Refined the architecture; helped to clean and annotate the data
Te Taka Keegan	Provided guidance and critical feedback
David Bainbridge	Provided guidance and critical feedback
Mark Apperley	Provided guidance and critical feedback

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

Name	Signature	Date
Vithya Yogarajan		24/11/23
Jemma König		30/11/23
Te Taka Keegan		23/11/23
David Bainbridge		24/11/23
Mark Apperley		24/11/23

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Matauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 8: Analysing AIO Possession in Māori-Language Tweets (Unpublished manuscript)

Nature of contribution by PhD candidate	Sampled and annotated the data; conducted the analysis & created the visualisations; wrote most of the first draft
Extent of contribution by PhD candidate (%)	70

CO-AUTHORS

Name	Nature of Contribution
Andreea S. Calude	Provided guidance and critical feedback; helped annotate the data; wrote the introduction
Ray Harlow	Provided guidance and critical feedback; helped annotate the data
Te Taka Keegan	Provided guidance and critical feedback; wrote a personal reflection

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

Name	Signature	Date
Ray Harlow	R B Harlow	1-2-24
Andreea S. Calude		1-2-24
Te Taka Keegan		1-2-24

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Co-Authorship Form

Postgraduate Studies Office
Student and Academic Services Division
Wahanga Ratonga Mataurauranga Akonga
The University of Waikato
Private Bag 3105
Hamilton 3240, New Zealand
Phone +64 7 838 4439
Website: http://www.waikato.ac.nz/sasd/postgraduate/

This form is to accompany the submission of any PhD that contains research reported in published or unpublished co-authored work. **Please include one copy of this form for each co-authored work.** Completed forms should be included in your appendices for all the copies of your thesis submitted for examination and library deposit (including digital deposit).

Please indicate the chapter/section/pages of this thesis that are extracted from a co-authored work and give the title and publication details or details of submission of the co-authored work.

Chapter 9: When loanwords are not lone words: Using networks and hypergraphs to explore Māori loanwords in New Zealand English. International Journal of Corpus Linguistics, 28(4), 461-499.

Nature of contribution by PhD candidate	Extracted and annotated the data; conducted the analysis & created the visualisations; wrote most of the first draft
Extent of contribution by PhD candidate (%)	70

CO-AUTHORS

Name	Nature of Contribution
Andreea S. Calude	Helped annotate the data; wrote the background; added relevant literature
Te Taka Keegan	Provided guidance and critical feedback
Julia Falconer	Advised on statistical aspects of the analysis

Certification by Co-Authors

The undersigned hereby certify that:

- the above statement correctly reflects the nature and extent of the PhD candidate's contribution to this work, and the nature of the contribution of each of the co-authors; and

Name	Signature	Date
Andreea S. Calude		23/11/23
Te Taka Keegan		23/11/23
Julia Falconer		24/11/2023

Appendix B

Ethics Approval for MultiCat User Study (Chapter 5)

The University of Waikato
Private Bag 3105
Hamilton, New Zealand, 3240
0800 WAIKATO (924 528)

HECS Human Ethics Committee
Brett Langley
Telephone +64 77 838 4060
Hecs-ethics@waikato.ac.nz

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

8 December 2023

David Trye
Mark Apperley
David Bainbridge
Andreea Calude

Re: HECS Ethics Approval of Application HREC(HECS)2023#70 "Evaluating MultiCat: A Visualisation Technique for Multidimensional Categorical Data."

Dear David:

Thank you for submitting your amended application HREC(HECS)2023#70 for ethical approval.

We are pleased to provide formal approval for your project, including the following activities:

- Recruitment of 5 to 10 participants to perform a small user observation study and questionnaire that evaluates a new technique called MultiCat
- Studies will take approximately 20 minutes
- Audio recording and screen recording consent will be sought
- Recorded data will be anonymised
- The datasets used for the purposes of this study are publicly available and are used within the spirit that they have been released

Please contact the committee by email (hecs-ethics@waikato.ac.nz) if you wish to make changes to your project as it unfolds, quoting your application number with your future correspondence. Any minor changes or additions to the approved research activities can be handled outside the monthly application cycle.

We wish you all the best with your research.

Kind regards,

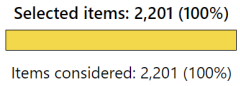
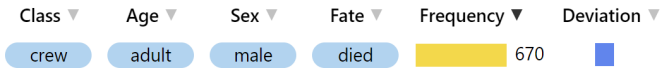
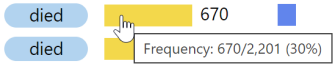
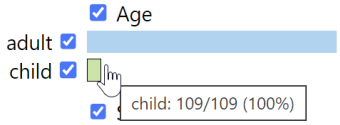
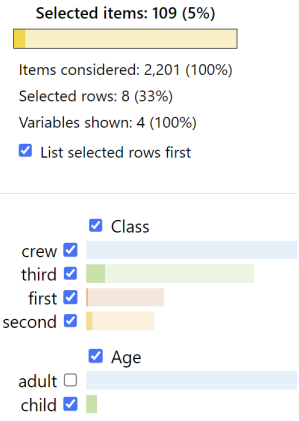
Brett Langley, PhD
Chairperson
HECS Human Ethics Committee
University of Waikato

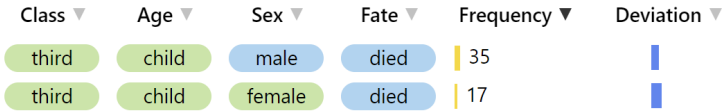
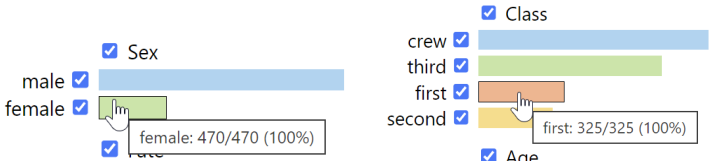
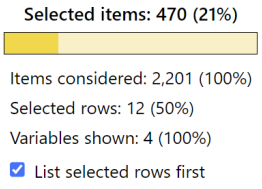
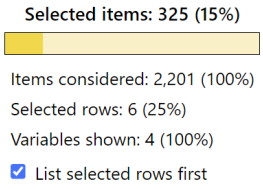
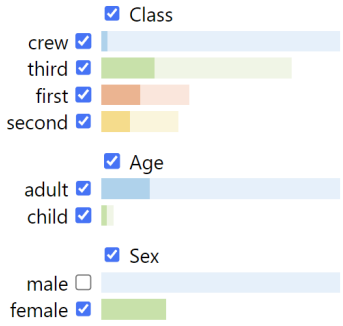
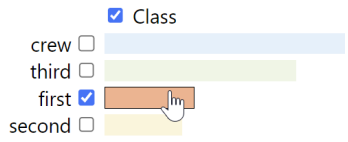
Appendix C

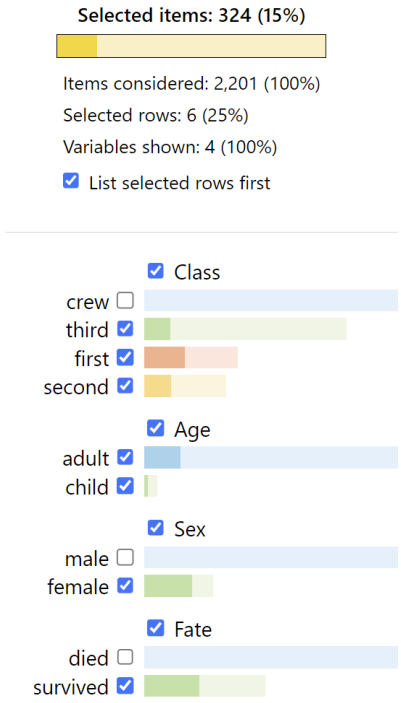
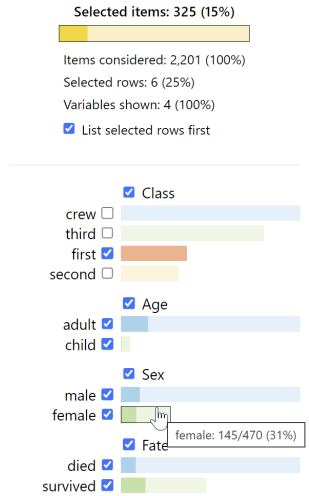
MultiCat User Study Tasks (Chapter 5)

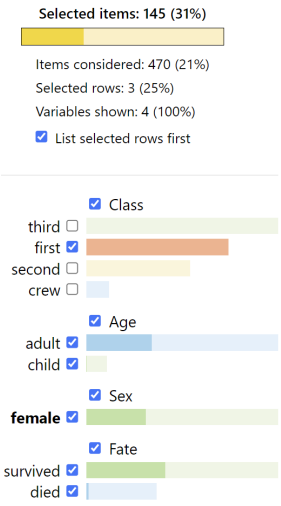
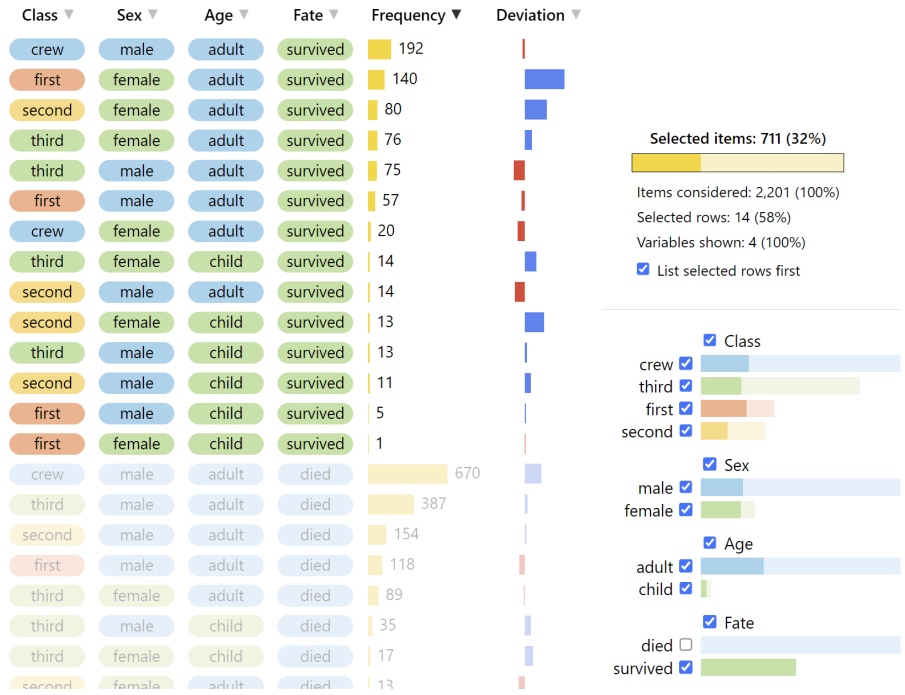
Supplemental Material: MultiCat User Study Tasks

Titanic Dataset (Dawson, 1995):

	Task Description	**Question (Q), Solution/Working (S) and Answer (A)**
T0	Summarise dataset	Q0: How many items (in this case, *people*) does the dataset contain? S: Read directly off the top right of the screen (top of the sidebar).  A: 2,201
T1	Identify key *N*-way relationship(s)	Q1a: What is the most frequent combination of categories (involving all four categorical variables) and how often does it occur? S: Look at the top-most row of the main visualisation.  A: {crew, adult, male, died}, 670 Q1b: What proportion of the total dataset does this combination account for? S: Hover over the combination's frequency bar to reveal the tooltip, then read the percentage.  A: 30%
T2	Find absolute value and marginal frequency for a particular category	Q2a: How many children were on board the Titanic? S1: Hover over the 'child' category in the sidebar.  S2: Select 'child', then look at the 'Selected items' bar chart.  S3 (inefficient): Remove all variables except 'Age', then read off the yellow frequency bar for children. A: 109

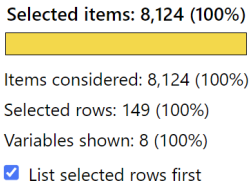
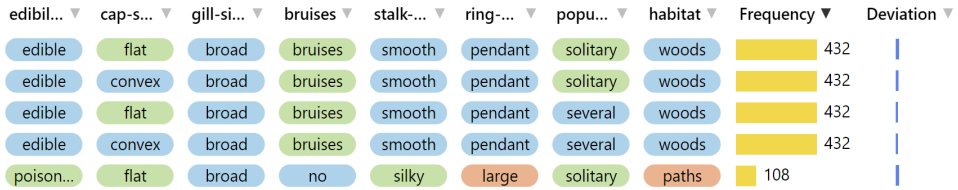
		Q2b: What percentage of the data do the children account for? S: After S2, read percentage (can't get answer directly from S1). A: 5% Q2c: What do you notice about the two most frequent combinations involving children? S: Look at the top two combinations (assuming data is still sorted by frequency) and identify columns where both stickers are the same. Participants may also comment on the deviation column (now called 'Residual').  A: They both relate to children in third class who died (boys first, then girls). Both combinations are slightly over-represented in the dataset.
T3	Compare frequencies of categories belonging to different variables	Q3: Which category is more frequent: 'female' or 'first' class? S1: Hover over tooltips for each category in the sidebar. *Don't* simply compare bar lengths as the bars for each variable are scaled independently (this is no longer the case; users *can* simply compare bar lengths).  S2: Select each category in turn and look at the 'Selected items' bar chart.     A: 'female' (470 > 325)
T4	Find non-conditional probability involving multiple categories	Q4: What proportion of people on board the Titanic were female passengers (i.e. non-crew) who survived? S: Select checkboxes for Class={first, second, third} (can just deselect 'crew'), Sex='female' and Fate='survived'. Read proportion from top right of screen.

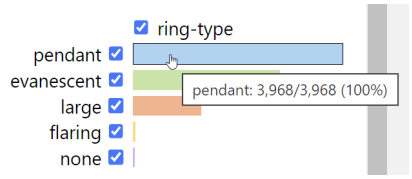
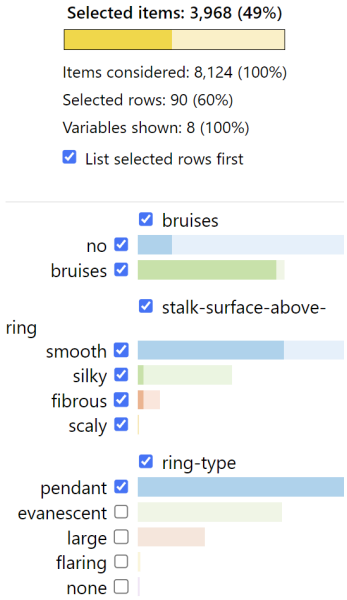
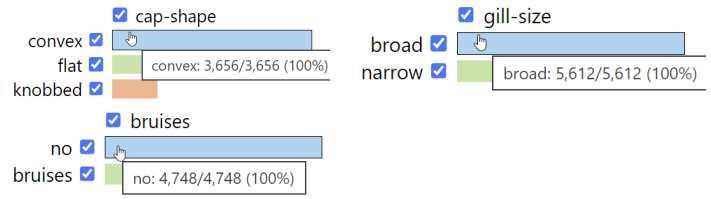
		 A: 15%
T5	Find conditional probability	Q5: What is the probability (as a percentage) that someone was in first class given that they were female? S1: Another way of phrasing this is *what percentage of females were in first class?* Select first class, then hover over females. 

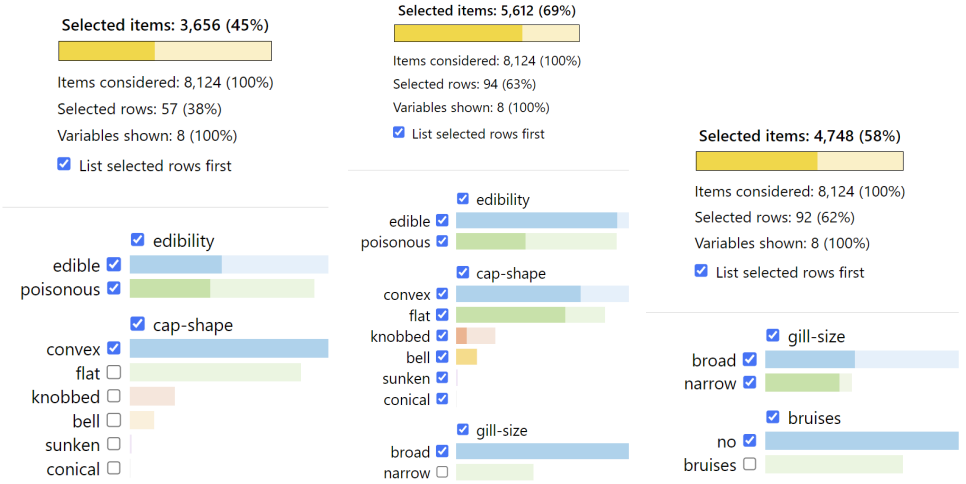
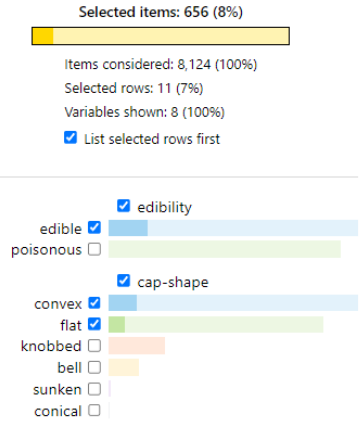
		S2: Select females, click ‘Filter by Selection, then select first class.  A: 145/470 = 31% (not 45% or 7%)
T6	Explore a (binary) response variable w.r.t. all other variables	Q6a: Let’s say you’re particularly interested in the people who *survived* the Titanic disaster. Do you notice any trends among this group of people? S: Select Fate=‘survived’, then examine selected combinations, as well as the proportion of selected data in the sidebar. Might choose to Filter by Selection (can then see, for instance, that 70% of survivors were passengers (first, second or third class), 30% were crew).  The most frequent combinations involve adults. Female class combinations are usually more frequent than corresponding male class combinations (two exceptions being adult crew, which is the most frequent combination, and first-class children). Class is mixed: no obvious trends, but can see female adults are ordered by first, second, third, while children are the opposite (for both sexes), presumably because there were not many children in higher classes. Looking at the sidebar: while a similar *number* of males and females survived, a far greater *proportion* of males died.

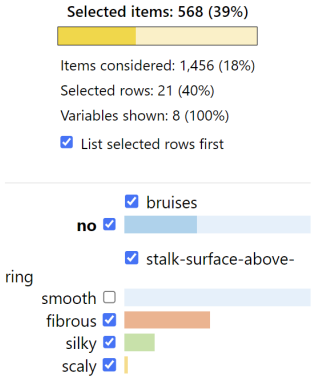
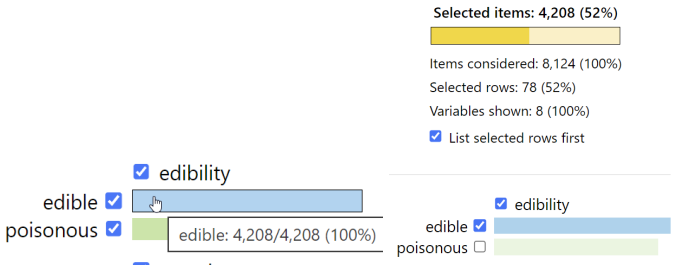
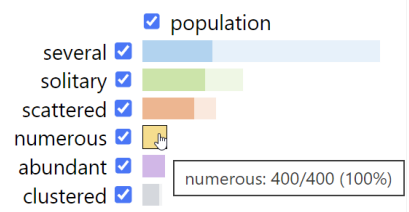
		Q6b: Do you notice any trends among survivors with respect to the deviations (over/under-represented groups)? S: Sort by deviation or manually scan largest values. A: The most over-represented combinations involve female survivors who were passengers (non-crew); conversely, the most under-represented combinations are male adult survivors.

Mushroom Dataset (Condensed): *The order of these questions was randomised.*

	Task Description	**Question (Q), Solution (S) and Answer (A)**
T0	Summarise dataset	Q0a: How many items (in this case, *mushrooms*) does the dataset contain? S: Look at the top right of the screen (top of the sidebar)  A: 8124 Q0b: How many categorical variables does the dataset contain? S: Look at the “Variables shown” metric. A: 8
T1	Identify key *N*-way relationship(s)	Q1a: How often do the most frequent combinations of categories occur? S: Look at the first few rows of the main visualisation (assuming it is still sorted by descending frequency). A: 432  Q1b: How many combinations with this frequency are there? S: Count the number of combinations whose frequency is 432. A: 4 Q1c: Do they share any of the same characteristics? If so, what are they? S: Look for same-coloured stickers in each column for all four combinations. A: Yes, all four are ‘edible’, ‘broad’, have ‘bruises’, ‘smooth’ stalks, ‘pendant’ rings, and are in ‘woods’ (only variables where categories differ are cap-shape, which is either ‘flex’ or ‘convex’, and population, which is ‘solitary’ or ‘several’)
T2	Find absolute value and marginal frequency for a particular category	Q2a: How many mushrooms have a pendant ring-type? S1: Hover over the ‘pendant’ category in the sidebar.

		 S2: Select ‘pendant’, then look at the ‘Selected items’ bar chart.  A: 3968 Q2b: What percentage of the data do they account for? S: After S2, read percentage (can’t get answer directly from S1). A: 49%
T3	Compare frequencies of categories belonging to different variables	Q3: Which category is the *least* frequent out of convex (cap-shape), broad (gill-size) and *no* bruises (bruises)? S1: Hover over tooltips for each category in the sidebar. *Don’t* simply compare bar lengths as the bars for each variable are scaled independently. 

		S2: Select each category in turn and look at the ‘Selected items’ bar chart.  A: convex (3656 vs 5612 vs 4748)
T4	Find non-conditional probability involving multiple categories	Q: What proportion of mushrooms are edible, have a convex or flat cap, and reside in scattered populations? S: Select checkboxes for ‘edible’, cap-shape={convex, flat} and population=‘scattered’. Read proportion from top right of screen.  A: 8% (656)

T5	Find conditional probability	Q5: What is the probability (as a percentage) that a mushroom does not have a smooth stalk surface, given that it is edible and has no bruises? S: Another way of phrasing this is *what percentage of edible mushrooms with no bruises did not have a smooth stalk surface?* We can't use the S1 approach from the Titanic dataset above as we want 'not smooth' and there is no way of hovering over a single merged category representing all other categories. As such, we have to select 'edible', 'no', then 'Filter by selection', then select all but 'smooth' for stalk surface.  A: 39% (568)
T6	Explore a (binary) response variable w.r.t. all other variables	Q6a: Let's say you're particularly interested in *edible* mushrooms (and you want to avoid the poisonous ones). How many edible mushrooms are there? S1: Hover over 'edible' category. S2: Select 'edible' category, then look at "Selected items".  A: 4208 Q6b: For which categories/properties can you be certain that a mushroom will be edible rather than poisonous? S: It is not (currently) possible to isolate these categories with a single query. Don't 'Filter by selection' as this means you lose sight of categories that overlap with poisonous mushrooms. Instead, select poisonous mushrooms and look for category bars that are fully opaque, meaning 100% of the category is selected; users can hover over each category in turn to ascertain whether this is the case, which is helpful for smaller bars (in updated prototype, the '100% bars' radio button is useful here)  A: sunken, flaring, numerous, abundant, waste

Appendix D

Metadata in the RMT Corpus (Chapter 6)

Table D.1: RMT Corpus V1 Metadata.

Variable	Type	Description
id	–	Twitter's unique identifier for the tweet.
content	Text	The tweet text, minus special characters.
content_with_emojis	Text	The tweet text, including special characters.
conversation_id	Cat.	ID of the tweet that initiated the conversation.
in_reply_to _user_id	Cat.	If the tweet is written in reply to another, this is the ID of the user who wrote the original tweet.
date	Temp.	The timestamp when the tweet was posted.
error	Cat.	The reason why the tweet could not be downloaded, if there was an error ('Authorization Error', 'Not Found Error', 'None').
favourites	Quant.	The number of favourites (likes, retweets & quotes) that the tweet received.
like_count	Quant.	The number of likes that the tweet received.
quote_count	Quant.	The number of times the tweet was quoted.
reply_count	Quant.	The number of replies the tweet received.
retweet_count	Quant.	The number of retweets the tweet received.
lang	Cat.	The two-letter code representing the language that the tweet was (erroneously) classified as.
media	-	Links to any photos or videos in the tweet.
outlinks	-	Any hyperlinks mentioned in the tweet.
source_label	Cat.	The device or third-party application from which the tweet was sent.
url	-	The link for viewing the tweet in context.
year	Quant.	The year the tweet was written (2007–2020).

Continued on next page

Table D.1 – RMT Corpus V1 Metadata (Continued).

Variable	Type	Description
maori_words	Text	The list of Māori words detected in the tweet.
num_maori_words	Quant.	The number of Māori words detected.
percent_maori	Quant.	The percentage of Māori text detected.
total_words	Quant.	The total number of words in the tweet.
user.id	-	Twitter's unique identifier for the user who wrote the tweet.
user.username	-	The username of the account.
user.alias	-	An alias for the author of the tweet in the form $T\langle X\rangle$, where $\langle X\rangle$ represents the user's ranking based on their total number of tweets in the corpus (user.num_tweets).
user.display_name	-	The user's display name on Twitter.
user.location	Text	The (unedited) location of the user.
user.region	Cat.	The user's location, based on self-reported information. Where possible, the data has been aggregated into New Zealand regions and names of overseas countries.
user.gender	Cat.	The user's gender if the account represents an individual ('male', 'female', 'gender-neutral', 'unknown') or 'group' if the account represents multiple people.
user.ethnicity	Cat.	The user's ethnicity, extracted from the account description (user.description).
user.iwi	Cat.	The user's tribal affiliation(s) if they are of Māori descent.
user.created	Temp.	The date the user's account was created.
user.description	Text	The user's self-written account description.
user.description_urls	-	Any links mentioned in the user's account description (user.description).
user.status	Cat.	The account status (as of December 2020) of the user who wrote the tweet: 'active', 'protected', 'suspended' or 'not found'.
user.favourites_count	Quant.	The total number of favourites that the user has received (not just counting tweets in the RMT Corpus).
user.followers_count	Quant.	The user's number of followers (as of December 2020).
user.friends_count	Quant.	The number of accounts that the user follows (as of December 2020).
user.link	-	The link for viewing the user's profile.

Continued on next page

Table D.1 – RMT Corpus V1 Metadata (Continued).

Variable	**Type**	**Description**
user.link_tcourl	-	The short version of any links associated with the user's profile (separate from their account description).
user.link_url	-	Any links associated with the user's profile (separate from their account description).
user.listed_count	Quant.	The number of people who have tagged the user in one or more tweets (not just counting tweets in the RMT Corpus).
user.media_count	Quant.	The number of tweets posted by the user that have included media items (not just counting tweets in the RMT Corpus).
user.num_tweets	Quant.	The total number of tweets in the RMT Corpus that were written by this user.
user.prof_banner_url	-	The link to the user's profile banner.
user.prof_image_url	-	The link to the user's profile picture.
user.statuses_count	Quant.	The total number of statuses that the user has posted (not just counting tweets in the RMT Corpus).
user.verified	Cat.	Whether or not the user's account is verified ('true' or 'false').

Appendix E

Algorithms for the Hybrid Architecture (Chapter 7)

Algorithm E.1 Token-Level Labelling

```
Input: Pre-processed tweets, list of Māori labels obtained from RMT system, pre-trained ML model, and tokenizer
Output: Labels at token-level
class_label = [ML model output]
english_list = [tokens with class_label 'E']
maori_list = [tokens with class_label 'M']
rmt_list = [Māori tokens from RMT system]
ambiguous_list = [rmt_list ∩ english_list]
if len(ambiguous_list) != 0 then
    Remove ambiguous tokens from rmt_list & english_list
end if
for each tweet i do
    for each token j in i do
        if j in english_list then
            if j is detected as an English word using fastText and NLTK language detection tools then
                Assign label for j as E (English)
            end if
        else if j in rmt_list then
            if j in maori_list then
                Assign label for j as M (Māori)
            end if
        else if j in ambiguous_list then
            Assign label for j as A (Ambiguous)
        else if Token j not in 'E', 'M', 'A' then
            Assign label for j as U (Unknown)
        end if
    end for
end for
```

Algorithm E.2 Context-Check for Ambiguous Items

```
Input: Pre-processed tweet tokens, list of Māori tokens, English tokens, and Ambiguous tokens obtained from token-level labelling
Output: Updated labels at token-level
for each tweet t do
    maori_list = [Māori words in t]
    english_list = [English words in t]
    ambiguous_list = [Ambiguous words in t]
    tokens = [all tokens in t]
    if len(ambiguous_list) != 0 then
        for amb_token in ambiguous_list do
            if amb_token contains {ā,ē,ī,ō,ū} then
                Assign label as M (Māori)
                Remove from ambiguous_list
            else
                before = tokens[index-1]
                after = tokens[index+1]
                before_before = tokens[index-2]
                after_after = tokens[index+2]
                if before & after in maori_list then
                    Assign label as M (Māori)
                    Remove from ambiguous_list
                else if before & after in english_list then
                    Assign label as E (English)
                    Remove from ambiguous_list
                else if before is null, i.e., amb_token is the first token in the tweet then
                    if after & after_after in maori_list then
                        Assign label as M (Māori)
                        Remove from ambiguous_list
                    else if after & after_after in english_list then
                        Assign label as E (English)
                        Remove from ambiguous_list
                    end if
                else if after is null, i.e., amb_token is the last token in the tweet then
                    if before_before & before in maori_list then
                        Assign label as M (Māori)
                        Remove from ambiguous_list
                    else if before_before & before in english_list then
                        Assign label as E (English)
                        Remove from ambiguous_list
                    end if
                end if
            end if
        end for
    end if
end for
```

Algorithm E.3 Tweet-Level Labelling

Input: Bilingual tweets with token-level labels obtained using Algorithm E.1 and Algorithm E.2
Output: Labels at tweet-level
for each tweet t **do**
    maori_list = [Māori words in t]
    english_list = [English words in t]
    unknown_list = [Unknown words in t]
    ambiguous_list = [Ambiguous words in t]
    if len(maori_list) == 0 & len(unknown_list) == 0 & len(ambiguous_list) == 0 **then**
        tweet_label of t is **E (English)**
    else if len(english_list) == 0 & len(unknown_list) == 0 & len(ambiguous_list) == 0 **then**
        tweet_label of t is **M (Māori)**
    else if len(ambiguous_list) == 0 & len(unknown_list) == 0 **then**
        tweet_label of t is **B (Bilingual)**
    else
        tweet_label of t is **O (Other)**
    end if
end for
for each tweet t **do**
    label_ML = ML tweet-label for t
    if label_ML == tweet_label **then**
        Final tweet-level label for MET Corpus
    else
        Further investigation needed
    end if
end for

Appendix F

Semantic Criteria for Māori Possessive Phrases (Chapter 8)

Table S1: Semantic Categories for Possessa (PSSM) and Possessors (PSSR)

Label	Description	Common Examples
part	● Indicates **how much or many**; the 'part' in a part-whole relationship (excludes *units* and *body* parts) ● Includes quantifiers, non-body parts, cardinal numbers, ordinal numbers and fractions ● Implies a *partitive* relationship	● *tētahi* (one) ● *ētahi* (some) ● *te nuinga* (most) ● *te mutunga* (the end) ● *te pito* (the bottom)
human	● **One or more people**, including: personal names; kinship terms (see <*kin* and >=*kin* relationships); descriptive nouns (e.g., expert, champion, *taonga* when referring to a person); job titles and agent nouns (Prime Minister, writer, singer); *iwi* (tribes) and *hapū* (sub-tribes); divine or supernatural beings; personifications (e.g., Mr Google, some instances of *taniwha*) ● Excludes personal *pronouns*	● *te whānau* (the family) ● *te Kuini* (the Queen) ● *kaiwaiata* (a singer) ● *ngā tūpuna* (the ancestors) ● *Te Ātiawa* (tribe)
pronoun	● A **personal pronoun** (*au, ahau, koe, ia, tāua, māua, kōrua, rāua, tātou, mātou, koutou, rātou*), **demonstrative** (*tēnei, tēnā, tērā, ēnei, ēna, ērā*) or other pronoun	● *koutou* (you 3+) ● *mātou* (we, excluding you) ● *tēnei* (this) ● *tātou* (we, including you)
institution	● **Entities** with a specific purpose that involve **human interaction** and collaboration, including schools, businesses, political parties and government departments	● *te kura* (the school) ● *Te Wharekura* (The Māori high school) ● *te Paati Reipa* (the Labour Party)
representation	● **Symbolic, nominal or visual representations** created by humans to convey meaning, identity or recognition ● Points to something but has no meaning in and of itself ● Typically implies a *representation* relationship	● *te ingoa* (name) ● *ngā tohu* (symbols) ● *te takoto* (layout)
unit	● Discrete elements used for **measurement, classification or organisation** (excludes *time*) ● Includes anything used to 'quantify' abstract nouns and uncountable nouns (e.g., 'pearls' of wisdom)	● *te rā tuatahi* (the first day) ● *ngā whārangi* (the pages) ● *Ngā Toenga* (The Remnants)
time	● A **period of time** or specific day/week/month/year, etc.	● *te wiki* (the week) ● *te wā* (the time) ● *te marama* (the month)
knowledge	● The collective **body of knowledge** developed by humans over time, including technological advancements, scientific discoveries, philosophical concepts, etc.	● *te kaupapa* (the purpose/policy) ● *te karere* (the news) ● *te mātauranga* (the knowledge) ● *hangarau* (technology)

cognition	• The **mental processes** of an animate possessor, including thoughts, emotions, desires, perceptions and other aspects of emotional functioning	• *te aroha* (the love) • *te whakaaro* (the thought) • *te take* (the reason)
activity	• An **activity/event** (such as a wedding or festival) or action, including a **nominalisation** of a verb • When used as a possessum, typically implies a *descriptor*, *nom_agentive* or *nom_other* relationship	• *te mahi* (the deeds) • *ngā mihi* (acknowledgements) • *te kōrero* (the speech)
property	• A **characteristic** or feature of someone or something (e.g. size, age, colour), including abstract nouns derived from adjectives. • Typically implies a *feature* relationship • Includes *mauri* and *wairua*, which may be used as metaphysical properties or in a more abstract sense (e.g., *te wairua o te hui*, 'the spirit of the meeting')	• *te mana* (the authority) • *te pai* (the quality) • *te mauri* (the lifeforce) • *te kaha* (the strength)
portable_obj	• A **portable object** (tangible, movable thing) created by a human, including small tools and appliances; • Excludes the following items: clothing and grooming accessories (*body*); things that are symbolic in nature (*representation*); and transport (e.g., skateboard = *transport*)	• *te toki* (the adze) • *ngā kete* (the baskets) • *te pū* (the gun) • *te karaka* (the clock)
large_obj	• A **large, man-made object** such as machinery and furnishings • Excludes the following items: buildings (e.g., houses, home, marae, mall = *place*); and transport (e.g., truck = *transport*)	• *te moenga* (the bed) • *te tatau* (the door) • *te tāhuhu* (the ceiling) • *ngā tari* (the offices)
digital_artefact	• Intangible entities that are created, stored, processed or utilised within **digital environments** • Includes digital services and product names (e.g. *ChatGPT*, *Photoshop*)	• *te kiriata* (the video) • *te reo irirangi* (the radio station) • *te nama waea* (the phone number)
cultural_ foundation	• **Intrinsic parts of a culture** that are transmitted from one generation to another. They provide the framework within which *cultural_artistry* is developed, understood and appreciated	• *Te Reo* (the [Māori] language) • *te mita* (the dialect)
cultural_ artistry	• (Non-digital) creations that are the product of **deliberate human effort** to convey emotions, ideas, stories and cultural values through various forms of artistic expression	• *te waiata* (the song) • *te pukapuka* (the book) • *te konohete* (the concert)
place	• A physical **location**, large/immovable structure, place name, geographical feature within a place (e.g., mountain, forest, sea) or location word (*māuī, matau, katau, mua, muri, raro, roto, runga, waho*)	• *te whare* (the house) • *te marae* (the meeting house) • *te taha* (beside/next to) • *te whenua* (the land)
transport	• Inanimate modes of **transport**, such as cars, bicycles and skateboards	• *te waka* (the canoe) • *tēnei pahi* (this bus) • *te motokā* (the car)

consumable	• **Food, drink,** medicine and cigarettes • Note that consumables are *a*-possessed, except for *wai* (water), *rongoā* (medicine) and *hikareti* (cigarettes), which are usually *o*-possessed	• *te kai (*the food) • *te keke* (the cake) • *kaimoana* (the seafood) • *kāwhe wera* (hot coffee)
flora_fauna	• **Flora and fauna**, including wild and domesticated animals, (types of) plants and other living organisms. • Most of this category is *a*-possessed, apart from animals used for conveyance, such as horses, which are *o*-possessed	• *ngā manu* (the birds) • *te kurī* (the dog) • *te wao nui* (the forest) • *te taiao* (the environment)
body	• **Body parts**, clothing or grooming accessories, such as ornaments, combs and watches • Includes body parts used as metaphors (e.g. '*heart* of the city' to mean 'city centre') • Includes effluence of the body, such as tears and sweat	• *te mata* (the face) • *te ringa* (the hand) • *te ngākau* (the heart)
other	• Anything that does not neatly fit into the above categories (e.g. celestial objects, atmospheric phenomena)	• *te rangi* (the sky) • *Matariki* (Pleiades star cluster) • *te rā* (the sun) • *te kapua* (the cloud)

Table S2: Semantic Relationships (RELA)

Label & Exp. Marker	**Description**	**Examples (PSSM → RELA ← PSSR)**
<kin (expect *a*, apart from *uri*)	• A familial relationship in which a human possessum is **generationally junior** to a human possessor. • The possessum may include but is not limited to: child (*taitamaiti*) or children (*tamariki*); son, boy or nephew (*tama*); daughter or girl (*tamāhine*); first-born (*mātāmua*); baby (*pēpi*); niece or nephew (*irāmutu*); daughter-in-law or son-in-law (*hunaonga*); grandchild (*mokopuna*); descendant (*uri*, but note this is usually *o*-possessed); nuclear family, of which the possessor is the head (*whānau*)	• *ngā tamariki a Tangaroa* = the children of Tangaroa (human → <kin ← human) • *te pēpi hou a tōku tuakana* = the new baby of my older sibling (human → <kin ← human) • *ngā uri o Poumatangatanga* = the descendants of Poumatangatanga (human → <kin ← human)
<non-kin (expect *a*)	• A non-familial relationship in which a human possessum is **subordinate** to, or under the protection of, a human possessor. • The possessum may include but is not limited to: student, learner or apprentice (*tauira, ākonga*); worker (*kaimahi*); patient (*tūroro*)	• *te pononga a Te Atua* = the disciple of God (human → <non-kin ← human) • *Ngā toa a Tūmatauenga* = the champions of Tūmatauenga (human → <non-kin ← human)
>=kin (expect *o*, apart from *wahine* and *tāne*)	• A familial relationship in which a human possessum is **generationally senior or equal to** a human possessor. • The possessum may include but is not limited to: mother (*whaea*), father (*pāpā*) or parents (*mātua*); uncle or aunt (*matua kēkē*); mother-in-law or father-in-law (*poupou*); grandmother (*kuia*), grandfather (*koro*); elder (*kaumātua*); ancestor (*tupuna*); extended family (*whānau*); wife (*wahine*), husband (*tāne*), even though spouses are usually *a*-possessed (unless preceded by *hoa*); younger sibling or cousin of the same sex (*teina*); older sibling or cousin of the same sex (*tuakana*); sister or female cousin of a male (*tuahine*); brother or male cousin of a female (*tungāne*); sibling-in-law of the same sex (*taokete*)	• *te Māmā o tō Pāpā* = the mother of your dad (human → >=kin ← human) • *ētahi mātua o ngā rangatahi* = some parents of the youth (human → >=kin ← human) • *te whānau whānui o Doug Tamaki* = the extended family of Doug Tamaki (human → >=kin ← human)
>=non-kin (expect *o*)	• A non-familial relationship in which a human possessum is in a **superior position** to, or is in some way responsible for, a human possessor. • The possessum may include but is not limited to: chief, boss, landlord (*rangatira*); leader (*tumuaki*); teacher (*kaiako*); doctor (*tākuta*); nurse (*nāhi*); guide (*kaiārahi*)	• *te rangatira o Ngāti Hine* = the chief of Ngāti Hine (human → >=non-kin ← human) • *te kaiako o taku pōtiki* = the teacher of my youngest child (human → >=non-kin ← human) • *te māngai o ngā māngai* = the speaker of speakers (human → >=non-kin ← human)

creation (expect *a*)	• The possessum was **created** or **produced by** an animate possessor, without whom it would not exist • Includes inventions, artistic works, personal expressions, etc. • Unlike *ownership*, creation is inalienable (i.e., non-transferable) • Excludes nominalisations of verbs (*nom_agentive, nom_other*)	• *te pukapuka a Hēmi Kelly* = the book of (written by) Hēmi Kelly (cultural_art. → creation ← human) • *ngā kōrero a Julian* = the words of Julian (unit → creation ← human) • *ngā whakaaro a te tangata* = the thoughts of the people (cognition → creation ← human)
ownership (determine marker w.r.t. possessum category)	• A non-human possessum such as an object, building or animal that **belongs to** or is **controlled by** the possessor • Includes things acquired by an action (e.g., buying, catching) of the possessor • Unlike *creation*, ownership can be changed/alienated (e.g., by gift or sale)	• *te kai a te rangatira* = the food of the chief (consumable → own. ← human) • *te toki a te whānau* = the adze of the family (large_object → own. ← human) • *te whare o Ngāti Uenukukōpako* = the house of Ngāti Uenukukōpako (place → ownership ← human)
creation/ ownership	• Used when unable to distinguish between *creation* and *ownership* due to insufficient context	• *Ngā kete o Māma* = Mother's baskets (portable_obj → creat./own. ← human) • *ngā wawata o ōku mātua* = the dreams of my parents (activity → creation/own. ← human)
partitive (expect *o*)	• A **part-whole relationship** where the possessum is the 'part/portion' and the possessor is the 'whole' • The possessum by its very meaning must have a possessor (even though sometimes implied): you can't be a *mema* (member) without being a member of *something*	• *tētahi o ōku tūpuna* = one of my ancestors (part → partitive ← human) • *te rā tuatahi o te wiki* = the first day of the week (unit → partitive ← time) • *te tihi o te maunga* = the peak of the mountain (unit → partitive ← place)
descriptor (expect *o*)	• Situates the possessum in relation to the possessor; often a **specifying or qualifying role** • Unlike *partitive* relationships, the possessum may stand alone, but is specified, limited or defined by a possessor • Includes phrases where the possessor is the **subject matter**, such as the topic of a song or book • Includes **creato*r*** relationships where the possessum is an agent noun (e.g., creator, author, composer) and created the possessor (this is the inverse of a *creation* relationship)	• *ngā rangatira o Waikato* = the chiefs of Waikato (human → descriptor ← place) • *te kaupapa o te wiki* = the topic of the week (knowledge → descriptor ← time) • *ngā kaihanga o #TheCasketeers* = the creators of [TV show] *The Casketeers* (human → descriptor ← digital_artefact)

feature (expect *o*)	• The possessum is a feature or property of the possessor, either an **abstract noun** or noun derived from an **adjective** • Includes temporary feelings (e.g., the *anxiety* of the man) • Includes, for instance, "the love of X" when the possessor is the *object/recipient* of the love (i.e., the intended meaning is the love someone independent of X has for X; cf. *nom_other*) • Excludes *representation*	• *te mana o te whenua* = the authority of the land (property → feature ← place) • *Ngā hara o ngā hōia* = the mistakes of the soldiers (property → feature ← human) • *te aroha o Aotearoa* = the love of (for) New Zealand (cognition → feature ← place)
representation (expect *o*)	• The possessum is a **symbolic, nominal or visual representation** of something that conveys meaning, identity or recognition, such as a flag, icon or abbreviation	• *te ingoa o te kurī* = the name of the dog (representation → representation ← flora_fauna) • *taku whakaahua o tō maunga* = my picture of your mountain (cultural_artistry → representation ← place) • *te whakamāoritanga o te curried heihei* = the translation of curried chicken (unit → representation ← consumable)
nom_agentive (expect *a*)	• The possessor is a subject of the nominalisation of either a **canonical transitive** verb (e.g., *āwhina, here, kawe, patu, pupuhi, whāngai*, see Bauer et al. 1997, p. 13) or an **agentive intransitive** verb (e.g., *haere, hoki, oma, piki*, see Bauer et al. 1997, p. 19) • The possessum is usually (but not always) an *activity*	• *ngā mahi a ngā tūpuna* = the deeds of the ancestors (activity → nom_agentive ← human) • *te mihi a ōna kaimahi* = the farewell of his/her co-workers (activity → nom_agentive ← human) • *te patapatai a John Campbell* = the interview of John Campbell (activity → nom_agentive ← human)
nom_other (expect *o*)	• The possessor is a subject of the nominalisation of one of the following types of verbs: **non-agentive intransitive** verb (e.g., *nui, pai, ora, rite*, see Bauer et al. 1997, p. 19); (intransitive) **neuter** verb (e.g., *mahue, mau, mutu, oti, pau*, see Bauer et al. 1997, p. 14); (transitive) **experience** verb (e.g., *kite, maumahara, mōhio, pīrangi, rongo*, see Bauer et al. 1997, p. 13); **passive transitive** verb (e.g., *whainga, patunga*) or **implied verb** (e.g., "the love of the group", where the intended meaning is the love *the group* feels (the group's love), rather than the love others have for the group, cf. *feature*) • The possessum is usually (but not always) an *activity*	• *te tīmatanga o te wiki* = the beginning of the week (activity → nom_other ← time) • *te tōnga o te rā* = the setting of the sun (activity → nom_other ← other) • *tō whakakāhoretanga a te rerenga* = your negation of the sentence (activity → nom_other ← unit)

Appendix G

Aggregating Hypergraphs by Node Attributes (Chapter 9)

Poster Details

Trye, D., Apperley, M., & Bainbridge, D. (2022). Aggregating hypergraphs by node attributes. In Angelini, P., & von Hanxleden, R. (Eds.), *Graph Drawing and Network Visualization: 30th International Symposium, GD 2022, Tokyo, Japan, September 13–16, 2022, Revised Selected Papers* (Vol. 13764, pp. 487-489). Springer Nature. `https://doi.org/10.1007/978-3-031-22203-0`

Abstract

PAOHVis (Buono and Valdivia, 2022; Valdivia et al., 2021) displays hypergraphs (Berge, 1973; Fischer et al., 2021) in a matrix where rows represent nodes (dots) and columns represent hyperedges (vertical lines). We propose extensions to PAOHVis for leveraging repeated hyperedges in non-simple hypergraphs, and displaying multiple node attributes. This is accomplished through two aggregation functions: *count-based*, which targets low-level detail, and *binary*, for high-level overview. In doing so, we introduce a domain-agnostic framework for consolidating hypergraphs by one or more categorical node attributes.

Preliminary results indicate that these enhancements provide a clearer picture of overall patterns and distributions of hypergraph data. Consider Figure G.1, which illustrates the different aggregation levels applied to a fictitious co-authorship dataset. There are 12 nodes (people) and 17 hyperedges

(papers), 13 of which are distinct. Nodes are coloured and subsequently consolidated by the gender of the author. The legend summarises node/category frequencies for the respective hypergraph. Additional categorical node attributes (e.g., affiliation, position and field) can be displayed and aggregated at the same time, provided these are mapped to different visual channels (e.g., shape, outline and texture). Unless they are strongly correlated, aggregating a larger number of attributes greatly reduces the number of identical hyperedges, resulting in a less compact visualisation. Thus, it may be more fruitful to aggregate hypergraphs by each attribute in turn, rather than attempting to visualise all attributes at once.

Count-Based Aggregation. This kind of aggregation shows, for each hyperedge, the exact number of nodes per category. Hyperedges with the corresponding number of nodes in each category (e.g., all papers authored by exactly two men and one woman) are combined. The original size of each hyperedge is preserved and nodes are stacked as tightly as possible, from the top row downwards, in descending order of overall category frequency. This layout facilitates comparisons of hyperedge size, which can be difficult to assess in non-aggregated hypergraphs. The original nodes (people) can no longer be reliably identified, since the same node in a repeated hyperedge may represent a different person across separate instances.

Count-based aggregation is useful for tasks relating to category frequency and overall set size. The middle panel of Figure G.1. shows that all papers have between two and four authors, which was not so apparent in the non-aggregated chart (left panel), due to the different line lengths. It is also easier to see that papers tend to have more male than female authors, but that the paper with the most authors of the same gender is written by four women (and no men).

Binary Aggregation. Hypergraphs can be further aggregated by collapsing each category with multiple occurrences in a hyperedge into a single node. The bar chart then shows the number of hyperedges that contain at least one node from *precisely* the corresponding categories. While this has been partially implemented in PAOHVis, it is not currently possible to consolidate identical hyperedges, which is essential for obtaining a quick overview of hypergraphs that are very dense, especially since (certain) hyperedges are likely to elicit higher counts, given the smaller number of possible category combinations. If an attribute has more than two categories, the data can be aggregated even further, so that all hyperedges are *flattened* into pairwise combinations.

Binary aggregation helps analysts to see how many distinct categories tend

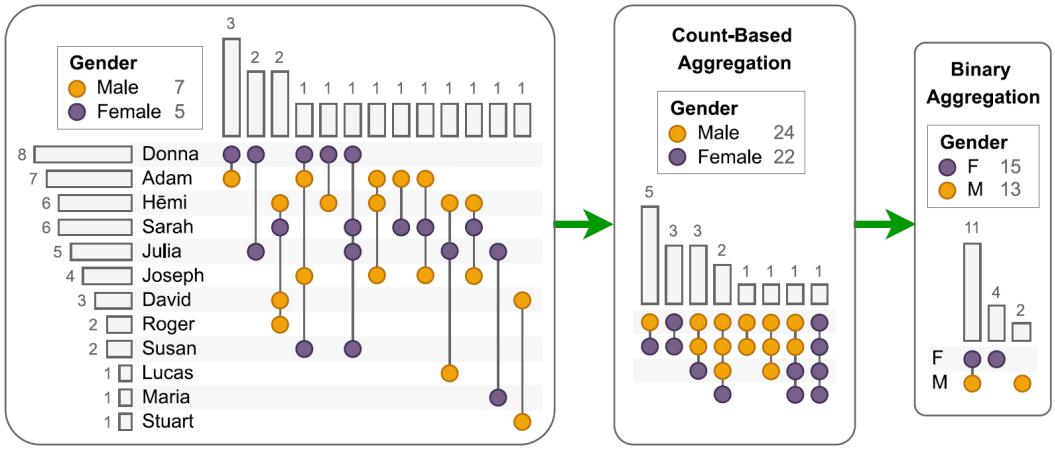


Figure G.1: Different levels of aggregation for a single node attribute (gender).

to occur in a hyperedge (e.g., do all categories occur together or only some?) and whether particular combinations of categories are dominant. The rightmost panel of Figure G.1 shows that, while papers tend to have more male authors, there are more papers authored solely by women (four) than by men (two).

In conclusion, building on PAOHVis, we advocate the consolidation of any repeated hyperedges and the encoding of their frequency in an aligned bar chart above each hyperedge. The result is visually similar to *UpSet* (Lex et al., 2014) but functionally different, with bar height denoting hyperedge multiplicity rather than set intersection size. This economises horizontal space, while also drawing attention to the distribution of recurrent hyperedges, especially when sorted by frequency.

Aggregation by node attributes is useful in situations where it is less important to know precisely which entities occur in relationships and more important to understand what *kinds* of entities they tend to be (e.g., to investigate a possible gender bias or to see how many papers have female-only or male-only authors). As the level of aggregation increases, more information about the original nodes and hyperedges is lost, in order to reveal more general patterns. It may be beneficial to view all levels of aggregation in conjunction, rather than in isolation.

References

Berge, C. (1973). Graphs and hypergraphs.

Buono, P. and Valdivia, P. (2022). Applications of dynamic hypergraph visualization. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, AVI 2022, New York, NY, USA. Association for Computing Machinery.

Fischer, M. T., Frings, A., Keim, D. A., and Seebacher, D. (2021). Towards a survey on static and dynamic hypergraph visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2(3):81–85.

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., and Pfister, H. (2014). Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992.

Valdivia, P., Buono, P., Plaisant, C., Dufournaud, N., and Fekete, J.-D. (2021). Analyzing dynamic hypergraphs with parallel aggregated ordered hypergraph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 27(1):1–13.

Appendix H

Productive Māori Loans in the Matariki Corpus (Chapter 9)

Table H.1: The 44 loans that 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 a song	MC	1	YES	2
pou	support poles	MC	1	NO	3

Continued on next page

Table H.1 – *Continued from previous page*

Loan	**English Counterpart**	**Semantic Domain**	**Size**	**Listed**	**Frequency Band**
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
...	the Māori language	PN	3	NO	2
Māori					
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

*As an outlier, *Māori* has been removed from parts of the analysis.

Appendix I

Māori Loans by Degree in the Matariki Corpus (Chapter 9)

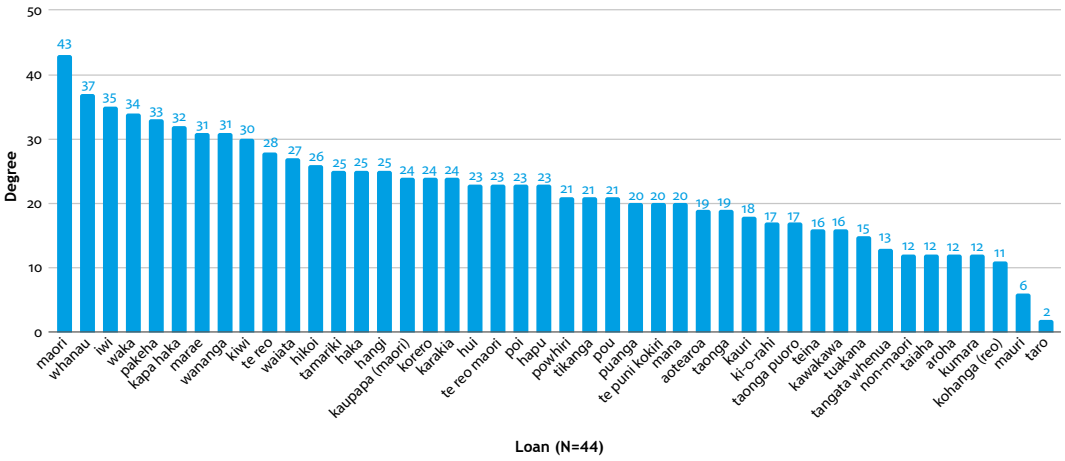


Figure I.1: Productive loans in the Matariki Corpus, ordered by number of connected nodes.

Appendix J

Sets including Māori in the Matariki Corpus (Chapter 9)

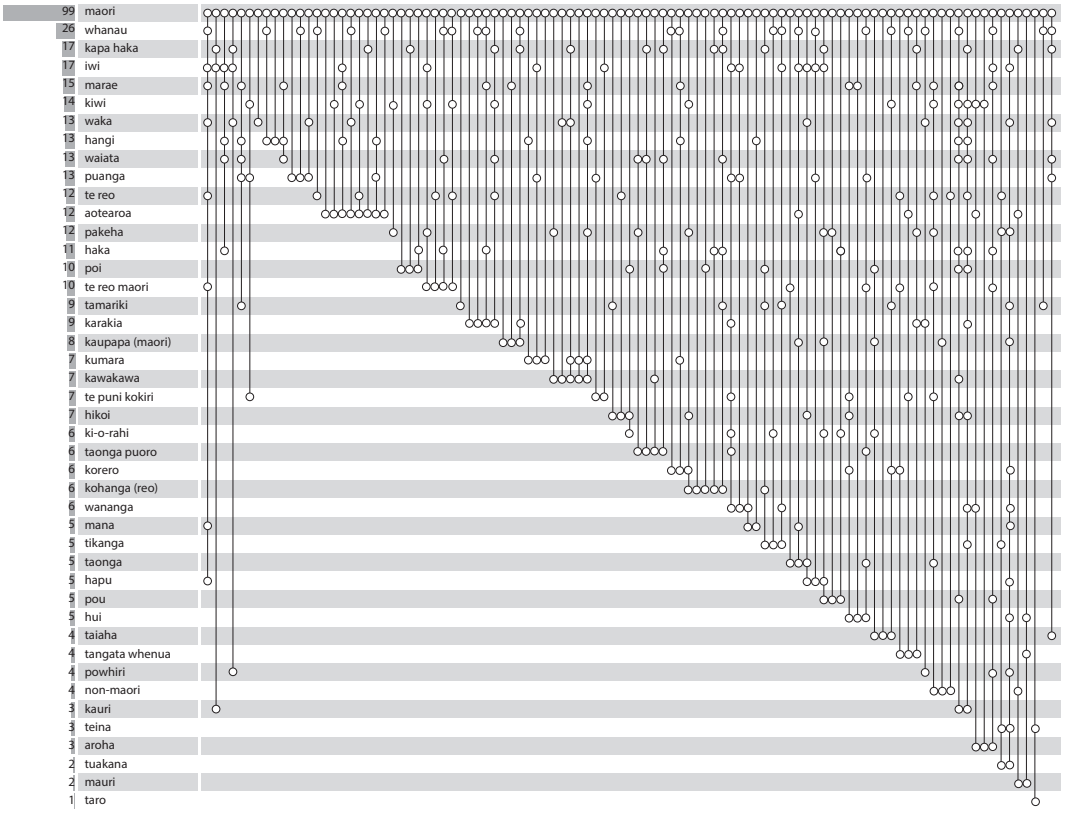


Figure J.1: All 125 sets in the Matariki Corpus, including the outlier *Māori*.