Paxing Tig outside the Crib
Dynamics of Lexical Variation in New Zealand English

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of Master
in Linguistics
by
Vicky Watson

University of Canterbury
2018
Contents

Abstract 5
Acknowledgements 6

1. Introduction 7
1.1 A History of Dialect Formation in New Zealand 7
1.2 Geographic & Demographic Breakdowns in New Zealand 9

2. Background 14
2.1 Regional Variation Worldwide 14
2.1.2 Use of Lexical Items as an Indicator of Regional Variation Worldwide 15
2.2 Variation in New Zealand English 17
2.3 Regional Variation in New Zealand 19
2.3.1 Attitudes towards Regional Variation 19
2.3.2 Phonological and Syntactic Regional Variation in New Zealand 22
2.3.3 Lexical Regional Variation in New Zealand 24
2.4 Aims, Hypotheses and Research Questions 26

3. Methodology 29
3.1 Background 29
3.2. “The Words We Use” 29
3.2.1 Questions 31
3.3 Overview of Collected Raw Data 33
3.4 Cleaning Data 35
3.4.1 Raw Data Decisions 35
3.4.2 Data Management 37
3.5 Analysis of Data 39
3.5.1 Classification Trees 39
3.6. Presentation of Data 40
3.6.1 ggplots 40

4. Results 42
4.1 “The place you go to watch a film” 42
4.1.1 Initial Observations 42
4.1.2 Age and Regional Distributions 43
4.1.3 Summary 45
4.2 “Another name for a holiday home” 47
4.2.1 Initial Observations 47
4.2.2 Age and Regional Distributions 47
4.2.3 Summary 50
4.3 “Item of clothing with long sleeves used to keep warm” 52
4.3.1 Initial Observations 52
4.3.2 Age and Regional Distributions 52
4.3.3 Summary 54
List of Figures

1. Map of New Zealand showing the 15 regions. Retrieved from Wikipedia (Reproduced with permission) ......................................................... 10
4. A figure from Chambers (2000:176) demonstrating the decline in regional influence over time, reproduced with permission ......................................................... 16
5. A figure from Chambers (2000:182) demonstrating the regional variation of “dresser”, “chest of drawers” and “bureau” in Canadian regions (reproduced with permission) ......................................................... 18
6. Example map from Neilson & Hay (2005:99) demonstrating a participant’s attitudes towards New Zealand regions (reproduced with permission) ......................................................... 21
7. Example of Bauer and Bauer (2003:10) results for the ‘chasing game’ (reproduced with permission) ......................................................... 27
8. Screenshot of the “The Words We Use” survey (reproduced with permission) ......................................................... 30
9. Visualisation of raw counts of participants ......................................................... 34
10. Proportional distribution of participants by age and region ......................................................... 36
11. Proportional distribution of participants by age group ......................................................... 37
12. Example of a classification tree ......................................................... 40
13. An example of a bean plot ......................................................... 41
14. Distribution of Q1 terms across age and island ......................................................... 43
15. Proportional distribution of film terms across region and other countries ......................................................... 44
16. Statistical distribution of the top four chasing game terms ......................................................... 45
17. Proportional distribution of top Q1 terms across region ......................................................... 46
18. Distribution of Q2 terms across age and island ......................................................... 48
19. Proportional distribution of second-house terms across region and other countries ......................................................... 49
20. Statistical distribution of the top two second-house terms ......................................................... 50
<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Proportional distribution of top Q2 terms across region</td>
<td>51</td>
</tr>
<tr>
<td>22</td>
<td>Distribution of Q3 terms across age and island</td>
<td>53</td>
</tr>
<tr>
<td>23</td>
<td>Proportional distribution of warm-top terms across region and other countries</td>
<td>53</td>
</tr>
<tr>
<td>24</td>
<td>Statistical distribution of the top two warm-top terms</td>
<td>54</td>
</tr>
<tr>
<td>25</td>
<td>Proportional distribution of top Q3 terms across region</td>
<td>55</td>
</tr>
<tr>
<td>26</td>
<td>Distribution of Q4 terms across age and island</td>
<td>58</td>
</tr>
<tr>
<td>27</td>
<td>Proportional distribution of chasing-game terms across region and other countries</td>
<td>59</td>
</tr>
<tr>
<td>28</td>
<td>Statistical distribution of the $t_g$ chasing-game terms</td>
<td>60</td>
</tr>
<tr>
<td>29</td>
<td>Statistical distribution of the top two chasing game terms</td>
<td>61</td>
</tr>
<tr>
<td>30</td>
<td>Proportional distribution of top Q4 terms across region</td>
<td>62</td>
</tr>
<tr>
<td>31</td>
<td>Distribution of Q5 terms across age and island</td>
<td>64</td>
</tr>
<tr>
<td>32</td>
<td>Proportional distribution of immunity terms across region and other countries</td>
<td>65</td>
</tr>
<tr>
<td>33</td>
<td>Statistical distribution of the top six immunity terms</td>
<td>65</td>
</tr>
<tr>
<td>34</td>
<td>Proportional distribution of top Q5 terms across region</td>
<td>66</td>
</tr>
<tr>
<td>35</td>
<td>Distribution of Q6 terms across age and island</td>
<td>69</td>
</tr>
<tr>
<td>36</td>
<td>Proportional distribution of truancy terms across region and other countries</td>
<td>69</td>
</tr>
<tr>
<td>37</td>
<td>Statistical distribution of top three truancy terms</td>
<td>70</td>
</tr>
<tr>
<td>38</td>
<td>Statistical distribution of top two truancy terms</td>
<td>71</td>
</tr>
<tr>
<td>39</td>
<td>Proportional distribution of top Q6 terms across region</td>
<td>72</td>
</tr>
<tr>
<td>40</td>
<td>Distribution of Q7 terms across age and island</td>
<td>74</td>
</tr>
<tr>
<td>41</td>
<td>Proportional distribution of no-friend terms across region and other countries</td>
<td>75</td>
</tr>
<tr>
<td>42</td>
<td>Statistical distribution of no-friend responses</td>
<td>76</td>
</tr>
<tr>
<td>43</td>
<td>Proportional distribution of top Q7 terms across region</td>
<td>77</td>
</tr>
</tbody>
</table>
Abstract

This thesis is a sociolinguistic study into lexical regional variation in New Zealand English. The study utilises qualitative responses provided in an online survey and analyses them using quantitative methods. Classification and Regression Trees (CARTs) are employed with fixed and random effects to determine statistically significant trends in the data. This study focuses on testing a number of social and geographical factors such as age, gender, mobility, and region in order to determine the influences over any given individual’s preferred lexical choice. Anecdotally, many people in New Zealand, whether linguistically trained or not, are aware that there is a difference between the way people from the Southland region speak in comparison to people who live in other areas of New Zealand. However, Bauer and Bauer (2000) propose a three-way split pertaining to the regional speech in New Zealand, suggesting a Northern New Zealand English as well as the well documented Southland New Zealand English. This means that there would also be a standard central New Zealand English spoken throughout the majority of the country. This thesis seeks to determine whether the lexical regional trends seen in the study of Bauer and Bauer’s participants are continued on into today. The online survey analysed in this thesis was available between 2012 and 2017 and contained 36 questions pertaining to the words individuals use to describe both everyday items, childhood games and daily interactions. For example:

- What do you call the public place you go to watch a film?
- If you missed school without permission, for example to go to the park or into town, what word would you use to describe it?
- A person who didn’t have any friends was called what?

Of the 36 survey questions, seven are analysed in depth in this study, using data from 2400 New Zealand-born respondents. Three questions regard everyday items, and tend to show strong influences of time, as well as hints of regional variation. The remaining four questions relate to childhood games, and show clearer regional patterns, as well as influences of time. Although not every question demonstrates a three-way split between the northern, central and southern areas of New Zealand, there is evidence that the northernmost regions behave differently to the rest of the country, in addition to the predicted divergent behaviour of the southern regions. The preferences differ for each question, with single lexical items being the majority variant for most respondents in some cases, and up to six lexical items being somewhat equally preferred terms in other questions. Overall, although change over time is significant in predicting lexical choice, there is also strong quantitative evidence to argue for regional variation country-wide. These results have implications for identifying and documenting regional variation in New Zealand English as it emerges in the vocabulary of New Zealanders.
Acknowledgements

A thesis is an undertaking that although at times may feel quite isolating is in fact the result of a whole team of people supporting the main researcher. This thesis in particular could never have been completed without a large number of people helping in their own particular manners.

My primary supervisor Kevin was a helpful sounding board, gently guiding me to uncover new research and inspiration for the project. He was also the one who created and distributed the survey. My secondary supervisor Vica was extraordinarily supportive and helpful with R code in particular, offering new ideas and proposing solutions that led to me making many discoveries regarding the capabilities of R, only a sample of which appear in the final version.

Regarding the NZILBB lab crew (esp Sidney, Hannah, Sarah, Nick, Jacq & Donald), I could never have finished this in the same time had I not had our tea and lunch breaks, discussing all manner of topics and helping each other battle R. A special shoutout to Sidney for reading over many iterations of this thesis, bringing snacks and baking into university, and being a wonderful sounding board for ideas and difficulties. I must also thank the members of the NZILBB who attended a Socio (brainstorming) meeting and gave me plenty of useful feedback about my topic.

Several academics were kind enough to allow me to reproduce images of their work. I hereby officially thank Dr Kevin Watson, and the Professors Jen Hay and Jack Chambers, as well as Emeritus Professor Laurie Bauer. Jack also wrote some very kind words of encouragement that came just at the right moment, and of course this thesis would not be the research it is without the pioneering work of Laurie and Winifred Bauer.

My family have also been unbelievably supportive, giving encouragement and commenting on results and various drafts. And finally, I must give a heartfelt thanks to my wonderful partner Jonathan, who kept me smiling and sane through the most difficult parts of the degree. He provided equal amounts of encouragement and distraction, and this thesis would be completed a lot later if it wasn’t for his support.

Thank you, all of you. To the unnamed people in this section who helped in other ways along this process, my apologies if I have missed you here. Know that I appreciate all of the help I received during this thesis.

Any mistakes remaining are, of course, my own.
1. Introduction

The existence of regional variation in New Zealand English is a relatively new area in the New Zealand linguistics field. New Zealand has only been an English-speaking country for just over 150 years. As “the newest native-speaker variety of English in existence” (Trudgill et al., 2000:300) it has taken some time for the process of dialect levelling to occur and for New Zealand English to form as a unique variety, distinct from the English spoken in the home countries of the 19th century settlers. This makes it a highly fruitful variety to explore for the simple reason that regional variation appears to only just be emerging. With different concentrations of a number of ethnicities within the two main New Zealand islands, it is perhaps no surprise that as New Zealand’s population increases, certain regions begin to develop identifiable speech features. Some scholars accept the idea that there may be no regional variation, yet many have observed that the Southland Region (the southernmost region in the country) has unique dialectal features, in particular Bartlett (2002). Additionally, a few studies have systematically tested for variation across a number of regions (Bauer & Bauer 2003; Ainsworth, 2004; Kennedy, 2006; Marsden, 2013).

In this thesis I will analyse a large dataset of lexical choices made by New Zealanders from all regions in the country. I will argue for the linguistic division of the country into three areas based on the responses given by New Zealanders born in those three areas. In this exploration of lexical variation I will also evaluate the effect time has had on the use of particular linguistic items. Chapter 2 will begin with a discussion of previous work into regional variation worldwide, and then lead into research into regional variation within New Zealand. In Chapter 3 the methodology shall be presented, as well as processes undertaken to complete the research questions regarding regional variation. Chapter 4 contains results, with comprehensive discussion appearing in Chapter 5. Finally, an overview is given of the limitations in the project in Chapter 6, as well as suggestions for future directions.

1.1 A History of Dialect Formation in New Zealand

Following the signing of the Treaty of Waitangi | te Tiriti o Waitangi in 1840 between the British Crown and various iwi, there was a great wave of immigrants arriving from the United Kingdom (UK). These immigrants would have spoken their own regional dialects.

Trudgill et al. (2000) presented the three stages of dialect formation they believe occurred in New Zealand. The first occurred between 1840 - 1860 when there was no “New Zealand English” as such. Instead, a variety of individuals from different social, cultural and national backgrounds were interacting face-to-face on a regular basis, where they would have no doubt been accommodating their speech in order to more easily communicate with their neighbours. These settlers came predominantly from Australia (which had been settled by English speakers from Britain in the early 19th Century), England, Scotland, and Ireland. The features that were deemed as less prestigious would likely have been lost in favour of more popular variants either on the ships, where the levelling process would have begun, or in the initial years of contact with unfamiliar dialects.
As noted in Trudgill, Gordon, Lewis, & Maclagan (2000), the second stage of dialect formation was a period of instability, which in New Zealand occurred approximately between 1860 and 1900. This is when the second generation of immigrants - the children - acquired features from the speakers around them (sometimes their peers, oftentimes their parents) and blended them into a new dialect of their own. Areas of New Zealand could be rather isolated and the children would have grown up in rather small communities, albeit sometimes with adults from very different backgrounds. Trudgill et al. (2000) noted that speakers born in New Zealand during this time period spoke with considerable inter- and intra-speaker variability. This variability then ‘focused’ into the stable, levelled variety. Although immigrants settled all over the country, there was an initial concentration in the South Island as a result of the gold rush, with a gradual influence of the less mountainous North Island as the excitement of the Gold Rush waned. The North Island had a more evenly distributed mixture of immigrants, with the socio-political influence of the North Island growing in the late 1800s. The reasons for this were twofold: transportation (and thus export) of cold products such as dairy became more possible, and the wealth coming from the Otago gold mines began to peter out (Schreier et al., 2003). There were some regions which became more densely populated with people from certain areas in the UK. The non-Māori, English speaking population of New Zealand increased almost six times over between 1860 and 1881 (a growth from 80,000 to 470,000)(Trudgill et al., 2000).

In 1871, Scottish immigrants represented between 60 - 80% of the total population of people living in the Otago Province (Olssen 1984:71 as cited in Schreier et al., 2003). Speech samples from both the Otago and the Southland regions in modern times have been found to demonstrate phonetic and lexical remnants of the Scottish English which many of the settlers in that area would have spoken (Bartlett, 1992; Bartlett, 2002). Canterbury, on the other hand, had a higher proportion of English settlers. As the Anglican religious settlement was in Canterbury, and much of Canterbury was settled by English people arriving from the more affluent southeast of England, there has long been an idea that Cantabrian (the demonym for a Canterbury resident) speech is somehow more English-like than that of other New Zealanders (Sinclair, 1991 and Gordon, 1997 as cited in Schreier et al., 2003). The West Coast had almost equal numbers of settlers from Ireland and England, although this seems not to have influenced the dialect of the West Coast as much as the Scottish settlement influenced Southland and Otago English. Having said that, a mere 14,860 people lived in the West Coast in 1871, in comparison to 85,113 inhabitants further south.

It was not until after World War I that New Zealand as a single nation, separate from “Mother Britain” began to form its identity, drawing the citizens together into one collective people. After the event that became known as ANZAC day (25 April 1915), when New Zealand and Australian troops suffered extraordinarily heavy casualties in Gallipoli, Turkey on behalf of Britain, New Zealanders began to see themselves as separate from British people. Although the radio and television presenters were encouraged to speak with a Received Pronunciation accent, which still resembled the dialect spoken by the British Broadcasting Company (BBC) in England, everyday New Zealanders’ speech was beginning to change. One famous example was a shift known as the Short Front Vowel shift, documented by Trudgill, Gordon, & Lewis (1998). At this stage, any regional variation would have been still levelling, as New Zealand English began to emerge as a dialect. Until then “regional variation” in New Zealand would
have been seen as a comparison between the English spoken in England compared to the English spoken in New Zealand (and Australia).

1.2 Geographic & Demographic Breakdowns in New Zealand

Before exploring regional variation, it is important to explain the geographic boundaries of New Zealand | Aotearoa and the proportions of its citizens in a modern context.

First of all, although New Zealand is an archipelago with many islands, there are two that the majority of the population live on, called respectively the North Island | Te Ika-a-Māui and the South Island | Te Waipounamu. In 2017 the population of New Zealand was 4,844,700 (Statistics New Zealand [StatsNZ], 2013). Within New Zealand, regional councils cover and govern all of New Zealand with the exception of the Chatham Islands. The government recognizes 16 regions and territorial authorities (with their own local government councils) in New Zealand: Northland, Auckland, Waikato, Bay of Plenty, Gisborne, Hawke’s Bay, Taranaki, Manawatu-Wanganui, Wellington, Tasman, Nelson, Marlborough, West Coast, Canterbury, Otago and Southland. The biggest region is Auckland with a population of 1,415,550 based on the 2013 Census from Statistics New Zealand (2013c) (StatsNZ, 2013) followed by Canterbury, with a population count of 539,433 as at the last census. New Zealand has 13 cities within these regions; meaning that not every region contains a city. Some regions such as Wellington contain multiple cities (Wellington City, Porirua, Lower Hutt, Upper Hutt). Figure 1, sourced from Wikipedia (2018) shows a colour coded map of New Zealand with its regions(Source: Wikipedia. Attribution 4.0 International licence).

There are four principal ethnicities: New Zealand European, also known as Pākehā, who make up 74% of New Zealand’s citizens and residents; Māori, indigenous Polynesians who settled in New Zealand around the 1300’s, who make up 15% of the population; New Zealand Asian (12%), the largest growing ethnic population in New Zealand, and New Zealand Pasifika (7%). These four groups account for the majority of the population. An actual percentage is unknown as individuals may identify with multiple ethnicities. Of the 598,602 people that identified with Māori ethnicity in 2013, 23% of those lived in Auckland region. Figure 2, drawn from Statistics New Zealand (2013a) shows the distribution of NZ European populations around the country (Source: Stats NZ and licensed by Stats NZ for reuse under the Creative Commons Attribution 4.0 International licence) whereas Figure 3, from Statistics New Zealand (2013b) shows the distribution of Māori populations around the country (Source: Stats NZ and licensed by Stats NZ for reuse under the Creative Commons Attribution 4.0 International licence).

Britain (2010) pointed out that regional boundaries are not only legally defined, existing on in a geographical plane, but that they also can be argued to exist on a psychological and social plane. Britain (2013) also mentioned that the ‘geographies’ of the elderly and the very young were far more constrained through reasons of money, authority and physical mobility than people aged between 20 - 60 years of age. Exploring regional variation cannot rely purely on the linguistic output of residents from a particular area, it should take factors such as age, gender, urbanization, network strength and socio-economic status into account. He
Figure 1: Map of New Zealand showing the 15 regions. Retrieved from Wikipedia (Reproduced with permission)
provided an example of a mobile class (in the United Kingdom) where middle-aged males would commute into London each day, away from where they lived. Additionally, these businessmen would have the financial means and the international networks to travel more frequently than the shopkeepers for instance of the same town who did not commute, or earn the same amounts of money, creating quite different linguistic patterns despite coming from the same “regional area”. In New Zealand regions and geographical areas could be grouped by transport routes (for example, there is no nation-wide or even island-wide railway system, most people travel either by car or by aeroplane).
2. Background

The ability of non-linguistically trained people to observe and distinguish dialects within a geographical domain has been a focus of dialectologists for a number of years. There has been ample work exploring how phonetic, syntactic and lexical differences are all important clues for identifying individuals through their speech. This chapter will explore how people have explored regional variation around the world and within New Zealand, as well as elaborate on where certain gaps lie in this area, particularly from a New Zealand English perspective. New Zealand English (NZE) and therefore variation in NZE did not exist at the time of the oldest studies (See Chapter 2: Section 1). When research into NZE variation began, social factors were very much considered and the methods will be covered in Chapter 2: Section 2.

2.1 Regional Variation Worldwide

This section will discuss research on regional variation and the perception thereof around the world and over time. People have always been interested in the way they spoke, and been aware that their neighbouring regions and countries can sound different to themselves. Marsden (2013) noted that the word ‘dialect’ was borrowed from Ancient Greek. The reason why English speakers refer to the Frenchman who conquered England in 1066 as William, yet the French call him Guillaume is due to regional variation occurring between the Norman French that the conquerors spoke in northern France, and the Parisian dialect that eventually became ‘standard’ French. Marsden also pointed out the historical langue d’oc/langue d’oil divide that existed in France until the 19th century in France. This referred to the fact that people from the northern and southern areas of the country had different dialects. The terms “oc” and “oil” were used to represent the common affirmative term “yes”, before it developed into today’s “oui” (2013).

The United Kingdom also notably has many regional dialects, evolving from the social and historical background of various invasions and mass migrations Watson (2018). Among others, Trudgill (2000) noted that in addition to social factors, prominent geographical features also play a significant role in the demarcation of dialect boundaries. For example, mountain ranges and large rivers could be considered prominent geographical features that inhibit movement. Britain (2013) addressed more societal features that could inhibit movement, such as particular public transport routes. Naturally, as transport and travel between areas becomes easier, more people are able to move longer distances and certain types of geographical features can become less of a hindrance.

Contemporary research investigating regional variation has been on-going since at least 1889 with The Existing Phonology of English Dialects, the first major undertaking of a regional variation project (as cited in Watson, 2018). This work mainly focused on the sound differences across the individual regions of England; therefore it was not until the English Dialect Dictionary, published by Wright between 1898 and 1905 that linguists really started probing into lexical variation. The data was collected using questionnaires and posted to localities around the United Kingdom. Since then, there have been a number of
dialectology projects undertaken with a focus on various linguistic aspects (phonological, syntactic, morphological and lexical) across the United Kingdom (Orton, 1962; Kolb, 1966; Orton & Wright, 1974; Orton, Sanderson & Widdowson, 1978 as cited in Watson, 2018). Social factors such as age, gender and class were not initially included in such investigations and Labov (1972), of course, was at the forefront of investigating the social influence on linguistic variables in the USA during the mid-late 20th century. In direct contrast to balancing social influences, field workers mainly looked to gather their data from Non-mobile, Older, Rural Males, or NORMS, who were assumed to speak the most rural and local ‘version’ of their dialect. Although some linguists do still use NORMS as the benchmark for how speakers from a particular area would speak, they are arguably no longer seen as the prototypical speakers of a language variety. This is due to the greater mobility and an increase in urbanization throughout the world.

Britain (2010) also discussed a period of time when geographers and linguists both began working on dialectology. He argued that in the light of the ‘quantitative analysis’ boom, there was valuable qualitative data being lost as individuals’ data became generalized over, and their responses became part of statistical predictions. It is important to maintain a qualitative aspect to regional variation, and find a middle ground where statistics and models can test the correlations of multiple factors. Yet it is equally important to maintain and consider data qualitatively with the idea that no two individuals have the same background.

2.1.2 Use of Lexical Items as an Indicator of Regional Variation Worldwide

In this section I outline some case studies undertaken in the USA and Canada which have examined lexical variation across age and regions.

As mentioned, with the increase in mobility, the influence of a region on an individual’s speech pattern decreases, as people are able to interact with speakers from other areas more easily and often. Chambers (2000) demonstrates in Figure 4 how in the 1930s region was the most helpful social variable in predicting a speaker’s linguistic response. However, by the 1990’s region dropped to the least influential social factor. Meanwhile, other social factors such as age, ethnicity and gender became more important predictors as well as education, although the latter showed the smallest change over time. Certainly, researchers examining data collected any time after the mid-20th century would be wise to incorporate, or at the very least account for these other social factors into their statistical models and analysis.

Chambers (1990) was one of the latest researchers investigating the lexical change over time in the USA and Canada for the word used to describe a ‘long piece of furniture’. As chesterfield was replaced in the USA by other terms during the 20th century, in Canada its use was growing. By the 1950’s some 75% of respondents in a survey said they used this term exclusively, whilst nearly nine in every ten people along the Canada-United States border employed chesterfield as their preferred term. In the 1970’s davenport was the most frequent lexical description for speakers of Midwest English in North America (inclusive of both Canada and the United States) (Allen, 1973-76 as cited in Chambers, 1995). Within this 1970’s study, the word chesterfield occurred five times. Each of those occurrences came
Figure 1: Statistically significant linguistic correlations with social variables in the 1930s and in 1990 (based on Johnson 1996: Table 6)

Figure 4: A figure from Chambers (2000:176) demonstrating the decline in regional influence over time, reproduced with permission
from a Canadian area. Although described as ‘uniquely Canadian’ in the Midwest, there was evidence for its use in California in a different study (Cassidy, 1985 as cited in Chambers, 1995). This shows changes over time, however Chambers (1990) was curious about how this change mapped onto the Canadian regions. Over 1991-1992 he undertook a postal survey in the “Golden Horseshoe” region, a geographical area encompassing a sixth of Canada’s population at the time (five million people). Chambers (1990) discovered that both age and region influenced the lexical item provided by any given participant, although age was the strongest factor. This suggests that when undertaking a study into lexical variation it would be essential to incorporate age as a variable, as the two interact.

In 2000, Chambers interviewed speakers from Quebec, Canada and questioned them about a variety of lexical terms. He added a regionality index, noting whether participants or their parents were born in the same region, whether participants still lived in that same region and whether or not they were raised as a child/teenager there. Participants were first and foremost placed according to where they lived during their formative years Chambers (2000). Participants were graded into one of nine categories on the index; a score between one and three indicated a good representative of the region, a score between four and five denoted a fair representative and a score between six and seven illustrated someone who was marked as a ‘linguistic interloper’ by Chambers. Preferences were clear, especially with an example as seen in Figure 5 regarding ‘the piece of furniture where you keep your socks, underwear, and other clothing’ (Chambers, 2000:182). Here it is clear that locality influenced the term a participant may select. Bureau was the name of choice for people who had been born and raised in Quebec, with a decline as the locality (represented by the regional index) decreases. A regionality index would likely be useful for future studies, as many people live in more than one place during their lifetime.

2.2 Variation in New Zealand English

Numerous studies have documented the development of New Zealand English over time, too many to note here; the following mentioned works are a subset of the research undertaken regarding the New Zealand accent. New Zealand English and its evolution has been documented in a wealth of literature therefore there will be little elaboration here. A more pertinent question at this stage is:

- How fully does change over time explain variation within New Zealand English?

Furthermore:

- How does change over time interact with other variables such as gender, region and social class?

There is a highly useful resource housed at the University of Canterbury called the Origins of New Zealand English Project (ONZE) which contains recordings of speakers born in New Zealand as early as 1851. Unfortunately for a researcher of regional variation, the recordings from the earliest periods of English-speaking New Zealand history did not take place across all
regions of New Zealand, only Otago/Southland, Wellington, Manawatu, Taranaki, Waikato and Northland.

As demonstrated in Labov (1972), gender can be an influential social factor in affecting dialects. Insofar as New Zealand English is concerned, a number of linguists such as Taylor (1996) and Fiasson (2015) have aptly demonstrated variation between the performance of female and male speakers. To date, little-to-no correlation has been found between gender and regional variation. There is currently some research investigating how word choice varies across the different genders (as well as social classes and ages) in New Zealand in casual speech regarding lexical variation interacting with gender (White & Hay, in preparation).

Ethnicity is always difficult to analyse, especially within the New Zealand context where every single person living in New Zealand today is or is descended from, immigrants. There are many linguistic studies exploring the differences between what Stubbe & Holmes (2000) calls “Pākehā English”, spoken by the majority of the country’s citizens and other Englishes spoken in New Zealand by New Zealanders that do not identify with Pākehā ethnicity, such as Māori or a subset of NZ Pasifika. For example, Szakay (2008) discovered that it was possible to identify an New Zealand individual’s ethnicity by their pitch. Bell & Gibson (2008) discussed phonetic differences between the speech of a variety of NZ Pasifika speakers. Similar to gender and age, there is little research on how ethnicity affects lexical choice. Stubbe & Holmes (2000) pointed out that there was a tendency for Māori speakers to use more Māori vocabulary when speaking English. There has however been increasing work regarding the variety of New Zealand English spoken in cities with large concentrations of
particular ethnicities. The “South Auckland Dialect” is fast becoming an area of interest and some related studies will be discussed below in Chapter 2: Section 3.2. What we can draw from this perhaps is regional variation could be influenced by larger or smaller concentrations of different ethnicities throughout New Zealand. Duhamel and Meyerhoff (2014:1) argue that due to ‘recent significant changes’ in New Zealand’s demographic distribution, it is highly plausible that new ethnic varieties could emerge, concentrated in certain areas of the country.

2.3 Regional Variation in New Zealand

The existence of regional variation within New Zealand English is a debated topic both between and within lay and linguistic circles. Trudgill et al. (2000:307) note the “remarkably small amount of regional variation” in New Zealand English. Vine (1999:14) was another researcher who stated that “NZE has little known regional variation”. Kuiper and Bell attribute a lack of regional variation to “the recency of migration and ... relatively free movement of settlers” (Kuiper & Bell, 2000:12). Another group of researchers have added a new angle to this perspective, noting that could perhaps be attitudes and stereotypes which fuel the convictions of distinctive regional variants (Bayard, 1990; Gordon, 1991; Holmes & Bell, 1990; Neilson & Hay, 2005; Duhamel et al., 2015). On the other side we see relatively few yet convincing studies uncovering pockets of phonetic and lexical regional variation around the country. Only two independent studies have investigated regional variation nationwide (Bauer and Bauer, 2003; Ainsworth, 2004). This section will evaluate both sides of the argument, paving a path towards perhaps a middle ground.

2.3.1 Attitudes towards Regional Variation

In this section I present an overview of perceptual dialectology before narrowing in on some New Zealand based attitude studies.

Dennis Preston was one of the pioneers in perceptual dialectology. His work was predominantly phonetic based and was centered around the dialects that occur in the United States (Preston 1989). His work has inspired the majority of the research regarding any regional variation within New Zealand (Bayard (1990); Neilson & Hay (2005)); Duhamel & Meyerhoff (2015)). Often speakers/participants are requested to rate speakers of suburbs, states, regions, and cities on traits and qualities such as intelligence, properness, friendliness etc. People have innate feelings about the aesthetic quality of particular accents. For example, some accents are said to sound pompous or posh while others are said to sound lazy and rough.

The question of regional variation in New Zealand is one that could be said up for debate. Evidence has shown that non-linguists certainly think there are perceptual distinctions between the regions. Anecdotally, many people have told me, on hearing my thesis topic, that ‘they always knew there were differences between regions’ and proceed to tell me about a time (or times) where they correctly deduced someone’s region of origin within New Zealand. New Zealanders and tourists alike seem certain that they can identify an Aucklander from a
Wellingtonian, or someone from the Hawkes Bay from the West Coast. P. Gordon (1997) tells an anecdote of how as a child she noticed the difference between the speech of her Otago-born father in comparison to her Northland-born neighbours.

In terms of investigating attitudes towards the regions, one of the earliest investigations into the ‘folk perceptions’ of regional New Zealand English was P Gordon (1997). She wanted to explore the divide between the perception of regional variation by lay people and the belief in regional homogeneity by linguists. All of the studies investigating perceptual dialectology use some form of replication of Preston (1989)’s study. 97 participants from the University of Otago responded to a questionnaire asking them to identify unique speaking regions of New Zealand. Of these, 95 (98%) believed there was variation within New Zealand English. 80% of the respondents pinpointed more than two distinct varieties. Many of the responses indicated influences of Māori accented English in the North Island (for example features such as High-Rising-Terminals (HRT), the discourse marker ‘eh’, and syllable timing) in comparison to a single mention in “South Island” English. Only people born in the Southland region reported discrimination due to their accent, although Gordon mentioned there were plenty of pejorative comments made about almost all of the regions with a sharp split between rural and urban areas. This study was small, however it provided a basis for future studies regarding the perception of regional variation.

Neilson & Hay (2005) conducted a similar experiment where 168 university students in Otago (37), Canterbury (42), Palmerston North (35) and Auckland (30) were given hand drawn maps of New Zealand and asked to write down their attitudes about people from the regions of New Zealand, as well as attitudes towards the speech of those regions/people. There were nine regions (some combined/condensed) given in this map: Northland, Auckland, Taranaki, Gisborne-Hawke’s Bay, Wellington, Nelson-Marlborough, Canterbury, Westland and Otago-Southland. Results showed that the regions with the larger cities (such as Wellington and Canterbury) were marked as more pleasant and correct than the more rural regions (such as Northland and the West Coast) with Auckland being the exception and being rated as correct but less pleasant than other well-populated regions. Figure 6 shows an example map filled in by a participant (Neilson & Hay, 2005:99) This study did not demonstrate the existence of regional variation, however, like Gordon’s study before it, it demonstrated that lay people are adamant that regional distinctions exist.

The research continued with yet another similar study, compiled by Duhamel & Meyerhoff (2015) at the University of Auckland, with a draw-a-map task, again in the fashion of Preston (1999). This study was completed a decade after the Neilson and Hay (2005) paper and challenged previous studies that the demographics within New Zealand were quickly changing, with a focus on the large (New Zealand) Asian and (New Zealand) Pasifika communities, particularly in Auckland. The predominant point made in this paper, however, was that certain commentaries about salient linguistic features in New Zealand English specifically mentioned ethnicity, social class and rural/urban-ness as driving factors behind the statements. The purpose was to find areas of New Zealand English in a regional context that could or would warrant further linguistic analysis. They discovered that people did not appear to group areas in New Zealand by regions necessarily, although 40% of participants identified Southland and Otago as a unique dialect area and 30% identified Northland. None of the lesser
Figure 6: Example map from Neilson & Hay (2005:99) demonstrating a participant’s attitudes towards New Zealand regions (reproduced with permission)
populated regions were mentioned in their article. Participants were also able to comment on specific Auckland suburbs. An urban/rural distinction was suggested as comments made by participants initially appeared to be split. However, participants were describing differences across a number of urban suburbs, so the urban/rural explanation was not fully satisfying. Another outcome was that the Auckland-based respondents seemed to be describing the dialects closest to them with social and ethnicity attributes, and the regions further away with linguistic attributes. However, the population and ethnic diversity is much larger in Auckland and Northland, and Southland was already well known for its ‘rolling r’. Having said that, participants were quite accurate when it came to describing the areas in Auckland where there were higher-than-average proportions of certain ethnicities.

There seems to be strong evidence for the attitude that regional variation exists. Therefore, if non-linguists believe they can identify accents from different regions, it would be reasonable to expect perhaps that there would be a level of accuracy of pinpointing regions when listening to New Zealand English speech from around the country. Watson and Watson (in prep) devised an experiment which asked participants to listen to clips and nominate where in New Zealand (or abroad) the speakers came from. Accompanied with the sound clips were Likert-scale questions designed to elicit attitudes towards the regions. 1002 responses were collected from 68 participants. Guessing the region had a success rate of 16%, and guessing the island had a success rate of 50%. This suggests that the identification was likely due to chance. People who sounded ‘posh’ to listeners were thought to be urban dwellers (predominantly Auckland or Christchurch). It seems ethnicity was also a guiding factor. Out of 68 listeners, all of them placed the segments with noticeable Māori-sounding accents in the North Island, despite the fact that two speakers were born and lived in Christchurch, New Zealand. Some participants indicated using lexical items as guides: “gigs don’t happen outside of Auckland” or an attempt to describe phonetic phenomena “Fellow pronounced fulla” or indeed sometimes a mixture “Has rural twang/ also uses upward inflections at the ends of sentences which is more north than south. Word heaps suggests more rural. but not 100%”. Without a confidence rating, it was difficult to surmise how many of the responses were guesses. Many of the comment boxes were left blank, others contained less helpful commentaries other than to reinforce the idea that regional variation seems to exist in the mind of non-linguists. “Doesn’t sound urban enough for Wellington”, “Sounds like a North Islander”, “Usual Christchurch accent”, “Sounds South Island but not deep South Island”.

These studies were, to reiterate, perception studies, with participant delineating the country themselves into ‘regions’ and annotating the maps with comments on how they believed inhabitants of each region spoke.

2.3.2 Phonological and Syntactic Regional Variation in New Zealand

The majority of research involving regional variation New Zealand English has involved exploring phonetic and phonological variation. The following will be a brief overview of some of the research investigating regional and phonetic variation. Regional variation was distinct in the study of Schreier, Gordon, Hay, & Maclagan (2003), who examined the speech of 45 New Zealanders born between 1896 and 1935 (collected in sociolinguistic interviews) from around
the country. These recordings were part of the Origins of New Zealand English (ONZE) database, which was created and maintained by the University of Canterbury and continues to be updated. The research question surrounded the /hw/ -> /w/ variability in words such as *what, whistle, somewhere* and *whether*. In this study, the significant factors found to be influencing the realization of the /hw/ sound were region, gender and age. Southlanders, women, and older speakers (born before 1914) were more likely to produce /hw/ than other speakers in the dataset. In terms of regional variation, speakers from Canterbury and the North Island behaved similarly, with 7.9% and 5.4% of /hw/ realizations respectively, in comparison to 40% of /hw/ realizations from speakers in the Southland region. This supports the evidence for a unique Southland accent, which has been well and truly demonstrated across many studies (see below). However, variation occurred between speakers from the Canterbury region and the North Island when it came to the type of words /hw/ was realized in. When the word was a content word (such as *whistle*), Cantabrians realized their words with /hw/ 50% of the time, in comparison to around 10% of North Islanders, and around 68% of Southlanders. Speakers from both Canterbury and the North Island rarely realized the /hw/ in function words, with a representative proportion of under 10% each.

By far and away, the most distinctive regional variant of New Zealand English is Southland English, with its famous ‘rolling R’ (a predisposition to realise the /r/ of the post-vocalic NURSE vowel). Bartlett (1992) investigated the rhoticity of speakers across the Southland region. Bartlett (2002) demonstrated several ways in which Southland English is a defined regional dialect with variation in phonology, grammar and lexis. In his thesis he listed 19 lexical items which he claimed were specifically associated to Southland English, one of them being *crib*, a word used to describe a second or holiday house. He describes this term as being used from Christchurch southwards (Trudgill & Hannah, 1982:21 as cited in Bartlett 2002). This Southland dialect is frequently the only one that ‘linguists unequivocally recognize’ (Bartlett, 2002:23). However, lexical analysis was not the focus of Bartlett’s thesis however. Bartlett made a point to specifically give his participants a closed list regarding lexical items. However, he did not comment on the items that people gave him in his analysis. From this absence of discussion one can assume that he did not find enormous lexical differences, although it could also be that participants responses all aligned either with each other and/or with the typical New Zealand English vocabulary. Any research investigating regional variation in New Zealand would reasonably expect to find Southland speech to have a number of different qualities to speech from other regions.

Variation within English spoken in the South Island seems to have been the focus of researchers in the 20th Century. In the current century, more light has been shed on variation in the North Island. Starks (2000) argued for a distinct difference in the way Aucklanders realised their /s/. Her focus was more on the ethnic variations of /s/ within the Auckland metropolitan area between New Zealand European, Pacific Island, Māori & Asian speakers who were all born in Auckland city or moved there before the age of five. Nonetheless, there is an argument for regional variation as there are higher proportions of Pacific Islanders and NZ Asians living in Auckland than in other areas of New Zealand. This fuels the debate over whether

---

1The NURSE vowel is part of the lexical sets designed by Wells (1982).
variation is in fact regional, or due to proportions of different ethnic populations, higher or lower socio-economic statuses or simply a split between rural and urban dialects. Essentially, one could argue that the proportional differences of ethnic populations across regions is what influences or creates the regional variation.

Kennedy (2006) looked at children’s speech, conducting 33 interviews at schools around New Zealand. Her focus was on the phonetic realization of six variable features of New Zealand English. Of these features, she attributed the variability of non-pre-vocalic /r/ and voicing of the final syllable segment in the word *with* to regional differences, specifically a difference between Southland and Otago realizations, and the realizations from speakers living in other areas of New Zealand. The other factors she claimed were the main influences of the remaining feature variability were ethnicity and socio-economic class. An important point of discussion was similar to Starks (2000): the uneven distribution of a variety of ethnicities within New Zealand, adding a layer of ambiguity to the driving influence of variation.

Marsden undertook a qualitative phonological study into two towns in Northern and Central New Zealand based on Bauer & Bauer, 2000, 2002a, 2002b, 2003’s studies. She interviewed teenagers from both towns and adults from the Northern region Marsden (2013). Her specific aim was to evaluate sociocultural factors and to see if the variation of linking /r/ in Māori New Zealand English was connected to speaker’s perception of regional identities in the Northern and central areas of the North Island in New Zealand. She generated a Māori integration index for each participant to uncover the strength of their Māori exposure and experiences. One argument to use teenagers is that they interact with primarily other people of their own age. In total she examined the speech of 51 adolescents and nine adults. Marsden discovered that a more rural/small town focus coupled with a high Māori integration index seemed to indicate less linking /r/ than for speakers that were New Zealand European or more urban focused.

Ainsworth (2004) discovered differences between speakers from rural Taranaki and Urban Wellington. She focused on the intonational variation of speakers and concluded that there were noticeable differences across the two areas that were influenced more by region than by factors of urban and ruralness. Another study looked at the difference between just two regions also. Additionally it is the only extensive study to date has involving syntactic regional variation in New Zealand English. Quinn (1995) examined the difference between the syntactic patterns of teenagers who lived in the West Coast and Canterbury regions of the South Island. She discovered significant differences across factors such as gender, socioeconomic status and region, as well as suggesting differences based on urban and rural populations.

### 2.3.3 Lexical Regional Variation in New Zealand

The largest work achieved in New Zealand regarding regional lexical variation was the “Playground Project” completed by L. Bauer & Bauer (2002a), L. Bauer & Bauer (2002b), L. Bauer & Bauer (2003) in the late 1990’s and analysed/published in the early 2000’s. The principal study involved creating linguistic quadrats where Bauer and Bauer collected one
sample (i.e. one response to the survey) per intermediate level school per 30 x 37 kilometres in New Zealand (150 schools in total). The reason they selected intermediate schools was to gather responses from participants old enough to give slightly more permanent and valid answers, but young enough to not be influenced by the teenage sphere of vocabulary. Teachers were asked to question their classes about certain terms (26 + questions) and note the answers. Some school provided multiple answers to the questionnaire while others just one. This study asked for phrases used at break times, friend-related terms such as greetings, farewells and names for items. Responses were stratified by region, island, whether the school was in an urban area or rural, whether the school was Catholic or not, and the decile rating of the school. Using the responses as guides, Bauer and Bauer divided the country into three principal regions and 11 sub-regions. These were approximately but not entirely representative of the territorial regions. They included partial regions in New Zealand and were described loosely such as Christchurch and most of Canterbury. Bauer and Bauer’s regions were as follows:

Table 1: Bauer and Bauer’s regional divisions

<table>
<thead>
<tr>
<th>Main region</th>
<th>Sub-region</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>East Northland</td>
</tr>
<tr>
<td>North</td>
<td>West Northland</td>
</tr>
<tr>
<td>North</td>
<td>Auckland</td>
</tr>
<tr>
<td>North</td>
<td>Central North Island</td>
</tr>
<tr>
<td>Central</td>
<td>Hawkes Bay - Wairarapa</td>
</tr>
<tr>
<td>Central</td>
<td>Wellington &amp; lower North Island</td>
</tr>
<tr>
<td>Central</td>
<td>Nelson - Marlborough</td>
</tr>
<tr>
<td>Central</td>
<td>West Coast</td>
</tr>
<tr>
<td>Central</td>
<td>Christchurch and most of Canterbury</td>
</tr>
<tr>
<td>Central</td>
<td>Timaru &amp; the Central Otago lake resorts</td>
</tr>
<tr>
<td>South</td>
<td>Southland and East and South Otago</td>
</tr>
</tbody>
</table>

These regions were somewhat unequal in size; thus a replication of this type of study would need to take geographic and demographic proportions into account. Additionally, this regional division shows linguistic, rather than governmentally defined borders, which makes further analysis slightly more difficult. Bauer and Bauer had very clear boundaries in their mind delineating the regions, however the descriptions are not sufficient to replicate with a different dataset from the outset. Bauer and Bauer (2003)’s regions were devised by running a statistical programme to test the levels of agreement of answers between each area. These ‘borders’ changed or disappeared depending on each question. For researchers looking to expand on or test the study, it could be useful to test the regional divisions using classification trees, to see if the data naturally grouped itself into recognizable regions/groups. Equally, generating some Classification and Regression Trees (CARTs) would lend itself to examining social aspects of the lexical terms also, and rank them in terms of influence. This means that it would be possible to see if gender, age, region and island correlated with each other. Despite this slight challenge, Bauer and Bauer’s three main regions have been used in studies by other researchers investigating regional variation within New Zealand (Ainsworth (2004),
Kennedy (2006)). One result demonstrating a three-way division was for the question related to a chasing game where one player has to run and try and touch another player. They found the majority response for the north, central and southern regions to be *tiggy*, *tag* and *tig* respectively. They also discovered specifically regional variation too. An example given was for truce terms (for example, when you’re playing a game and want to have temporary immunity whilst you attend to something such as tying your shoelace). Bauer and Bauer identified four regions that showed variation for this question: Auckland (*gates, quits*), Wellington (*fans, flicks*), Nelson-Marlborough (*taxes*) and Southland (*nibs*). Another regional linguistic difference Bauer and Bauer (2002) found was variation in the way students named someone who appeared to have no friends; people from Canterbury responded *nif*, which was short for *no identifiable/intelligent friends*. People from Auckland had another acronym *OTL*, which stood for *only the lonely*. When interpreting the results, Bauer & Bauer wondered whether the variation they found was age-graded, which would have meant that the childhood variation did not continue on into adulthood. A way to test this would be to ask adults of a variety of ages about which terms they used as a child. Bauer and Bauer (2003) claimed that although they found some evidence of island-divided lexical variation, most of their data fitted into one of the three regions listed above.

This study generated further data via media interest when a call went out to New Zealand Listener subscribers to send in the terms they had for certain household objects or childhood games, along with their age, gender, and location. In 2000, an article was published in *The NZ Listener* with Bauer and Bauer, requesting that listeners send in the terms that they used for two schoolyard terms. Over 600 responses were collected across the entire country (L. Bauer & Bauer (2002b)). Divisions such as the three way *tiggy/tag/tig* split were borne out. However, before the 1960’s, Bauer and Bauer discovered that although there was a three way split, it was more between the northern *chasey*, the central *chasing* and the southern *tig*. Similarly, the truce terms *nibs* and *fans* seemed to have their roots in the early-mid 20th century. One hypothesis posited by Bauer and Bauer (2002b) is that if lexical variation was influenced by the regions that settlers came from (i.e. southeast England, Denmark etc) then only the historical lexical items - words used by the settlers when they came - would demonstrate some regional pull. What this means is that we should not be able to link regional variation to the historical background of the settlers, unless the vocabulary was also historic.

### 2.4 Aims, Hypotheses and Research Questions

Building on previous research, this thesis sets out to be a modern nation-wide exploration into regional variation within New Zealand English. There appears to be a gap regarding the existence of regional differences within adult speech across the whole of New Zealand, where social factors such as age and gender can be accounted for. As Bauer and Bauer said:

>“Until we have looked for corresponding variation in adult vocabulary we cannot be sure that it [regional variation] does not exist”

(Bauer & Bauer, 2002a:181).

Although a single study (based on a request from a radio station) has initiated investigating
Figure 7: Example of Bauer and Bauer (2003:10) results for the ‘chasing game’ (reproduced with permission)
the variation within adult’s vocabulary in a regional context, this study will probe more deeply into the responses from thousands of adult New Zealand English speakers from around the country, using statistical software to test correlations between and across social and geographical factors. The overall aim of the thesis is to explore to what extent there is regional lexical variation in the production of New Zealand English adult speakers. As noted, several substantial investigations (L. Bauer & Bauer (2002b); Kennedy (2006); Marsden (2013)) have found a three-way division in New Zealand English between the north of the country (including the Northland, Auckland, Waikato and Bay of Plenty regions), the south of the country (comprising of the Southland, Otago and occasionally south Canterbury regions) and the ‘rest’ of the country, Central New Zealand. Therefore, I hypothesise that some three-way split between larger New Zealand areas will be apparent, particularly for the so called “Southland English”. If the forays into an emerging Auckland-based English are providing positive results, which they appear to be (Starks (2000)) then it would be hoped that participants of the current study coming from or based in and around the Auckland region would behave differently their more central and southern neighbours, especially among the results of the younger speakers. Given the ages of these regional divisions, it would be reasonable to assume that at least age (if not other social factors) would be significant contributors in predicting variation of speakers from around New Zealand. Additionally, Bell and Kuiper suggest that until recently, perhaps enough time has not passed in New Zealand history for regional variation to occur (2000:12). Now however, it is possible to compare one study on lexical regional variation with another, spanning nearly 20 years. It is hoped that if variation is age graded, then the results found in L. Bauer & Bauer (2003) are replicated for the respondents in this study who would have been 11 or 12 years old at the time of Bauer and Bauer’s study.

To summarise the hypotheses, I am expecting to find a regional division within the speakers of New Zealand English, all the clusterings are as yet somewhat uncertain. Additionally, I believe that age in particular will be a significantly influential factor in predicting the lexical choice from any given participant.

To test these hypotheses, I have three questions that I will answer over the course of this investigation:

1. Is there lexical variation in New Zealand between the two main islands?
2. Is there lexical variation in New Zealand between the regions?
3. Is it social factors such as age or gender, or regional factors that influence lexical variation around New Zealand?
3. Methodology

This chapter will cover the processes and justification measures taken to complete this project.

3.1 Background

This thesis is inspired by the work pioneered by Bauer and Bauer. Bauer and Bauer’s study was a novel approach to regional variation in New Zealand L. Bauer & Bauer (2003). The data they collected for their playground study was spread as evenly as they could manage (one school per 30 km * 37km squared) across the entire country. Only one response was collected per school per square. This means that Bauer and Bauer’s results, plotted onto a map of New Zealand as seen in Figure 7, were easy to distinguish across the regions. It would have taken too long to collect all of the data in order to replicate this study exactly. This method did indeed reveal some regional preferences, but it was unable to demonstrate density or popularity of tokens within one region. Fortunately, statistical analysis and technological capabilities has advanced considerably in 20 years, and I am now able to test Bauer & Bauer’s claims over a larger geographical space with more participants. The claims will be able to be robustly supported or challenged thanks to the use of advanced statistical software by the R Core Team (2017).

3.2. “The Words We Use”

The principal study makes use of an existing world-wide lexical survey named “The Words We Use”. It was designed to be a continuation of L. Bauer & Bauer (2003)’s ‘Playground Talk’ study, which began in 1999 in Wellington and has already been comprehensively covered elsewhere. Over five years between 2012 and 2017, the “Words We Use” survey was distributed electronically by the Linguistics Department at the University of Canterbury in Christchurch, New Zealand. The survey was designed and distributed by Dr Kevin Watson via the social media networks of his first year linguistics students. With Bauer and Bauer’s discoveries in mind, Watson’s primary aim was to provide first year undergraduate students with up to 500 survey responses with which they could conduct their own mini research projects around lexical variation in New Zealand English. For this reason, the dataset consists of raw, unstandardized responses collected over five years, which the students then standardized and analysed themselves for personal use (their assignments, rather than publication). Responses in the questionnaire were given in free text boxes, in order to give the most freedom and least amount of priming to participants. Figure 8 shows a screenshot of one page of the survey. The survey is still active and is added to each year. LING102 students request a particular topic (i.e. they wish to investigate gender or age-graded variation) and are provided with randomly selected observations in an R generated spreadsheet of raw data which they are expected to standardize and analyse from scratch. No-one to date has conducted analysis utilizing the entire dataset, nor has anyone performed this level of statistical analysis regarding the regional and age-graded variation.
The words we use

A: Words for things

A1. What do you call the public place you go to to watch a film?
   “I’m going to the _______ this evening”

A2. You have bought a mixed bag of things like caramels, wine-gums and toffees. How would you describe these?
   “I’ve got a bag of ________”

A3. You are going to the beach and want to swim in the sea, but you forget the clothes you would usually wear when you do this. What would you call the clothes in question?
   “I forgot my _________”
3.2.1 Questions

The survey contained 36 content questions such as *What do you call the public place you go to watch a film?*, split into four categories:

- A: general terms for items
- B: childhood games
- C: descriptions of feelings and states
- D: terms to do with friends and being social

For a full list of all the questions, please refer to *Appendix I*. Seven questions were selected to analyse in this thesis, three from the general section and four from the schoolyard section:

1. What do you call the public place you go to watch a film?
2. Do you have another name for a holiday home?
3. What is the name of the item of clothing, often woolen and with long sleeves, that you might wear over a t-shirt to keep warm?
4. When you were at school, did you play a game with many players where one player has to run and touch another player while all the other players try to run away and not get touched? If so, what was the game usually called?
5. Was there a word which you could say to show that you were not playing for a short time, for instance because you needed to tie up your shoelace?
6. If you missed school without permission, for example to go to the park or into town, what word would you use to describe it?
7. A person who didn’t have any friends was called what?

Three of these questions appeared in Bauer & Bauer (2003)’s study (4, 5 & 7). It would not have been possible in the given time frame to analyse all 36 questions. The first three questions were selected as they asked for names of terms which most adults would be familiar with and would be able to remember. The following four questions analysed were from the schoolyard section, selected because Bauer & Bauer previously found variation, and this would be a good point for comparison between studies. There were an additional eight demographic and administration questions in this survey, which asked after the participant’s age category, birth city, ethnicity and places they had lived for over three months of their lives.

In terms of regions, participants were asked their city of birth, and then this was manually coded for the following regions (sorted by island north to south and then latitude north to south). Additionally, locations were coded in a three way split in an attempt to align with the Bauer and Bauer (2002,2003) studies (henceforth referred to as “BB region”). The distribution of regions into island, BBregion and Legal region can be seen in *Table 3*, below.

<table>
<thead>
<tr>
<th>Island</th>
<th>BBregion</th>
<th>Legal region</th>
<th>Region number</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>North</td>
<td>Northland</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 2: Distribution of geographical areas
Longitude and latitudinal scores were also added for map making. They were added at
the region level, as not everybody provided a city. Each token was coded according to the
criteria shown in Table 3. The question names were named after the most frequent word in
each column. Longitudinal, latitudinal and regional category scores were added during the
analysis stage to facilitate plotting and modelling. Each region was assigned a number from
one to fifteen, to give an indication of how north or south a region was. These numbers can
be seen in Table 2. Although the Nelson-Tasman region is geographically more north than
Wellington, it was decided to keep the regions categorized within their own island first.

It was not possible to code for a regionality index as Chambers (2000) had done, however
participants were coded for having moved around in New Zealand during their formative
years, and whether they had lived outside of the country for more than three months. Due
to the high proportion of people coming from Canterbury, participants were also coded for
being born in the Canterbury region or not. Some people had lived in the same city all their
lives, but had travelled overseas, others had moved around New Zealand throughout their
childhood.

It was more difficult to standardize ethnicity, with 582 unique ethnicities given for people
born in New Zealand alone. Some of these could be collapsed (no-one would argue that
New Zealand European and European New Zealander were not by and large denoting the
same ethnicity) however many terms were accompanied by descriptions with strong opinions
attached, i.e. New Zealander NOT New Zealand European and Inappropriate question. Some
respondents seemed flippant, giving such answers as ‘ethnicity’, ‘what’s that lol’, ‘so white’,
‘awesome’, ‘my own’, ‘anglo-German Eurasian hybrid’ and ‘multiracial’. Yet another selection
of respondents provided multiple terms. In the end, it became just too difficult and time
consuming to mark each speaker with one or two ethnicities. An additional difficulty was the
overlap in terms. Kiwi and New Zealander denote nationality. There was a separate factor
coding for nationality, noting which country a person was born in. Both Māori and Pākehā
ethnicities easily fit into these groups, yet not all New Zealanders are Māori. Research has demonstrated linguistic differences between people who identify with either Pākehā, Māori or Mixed as described in Chapter 2: Sections 3, 3.1 and 3.2. Anecdotally and on social media, Pākehā is a strongly controversial term. The majority of respondents did identify with some form of New Zealand European/Pākehā identity, with 155 respondents identifying with some level of Māori ethnicity.

Table 3: Factors analysed in the survey

<table>
<thead>
<tr>
<th>Data vs Metadata</th>
<th>Column</th>
<th>Coded by the author or the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metadata</td>
<td>timestamp</td>
<td>survey</td>
</tr>
<tr>
<td>Metadata</td>
<td>participant number</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>gender</td>
<td>survey</td>
</tr>
<tr>
<td>Metadata</td>
<td>age category</td>
<td>survey</td>
</tr>
<tr>
<td>Metadata</td>
<td>age (1-7)</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>country</td>
<td>survey</td>
</tr>
<tr>
<td>Metadata</td>
<td>NZ y/n</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>move NZ y/n</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Metadata</td>
<td>move Overseas y/n</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Metadata</td>
<td>NZ island</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>city of birth</td>
<td>survey</td>
</tr>
<tr>
<td>Metadata</td>
<td>NZ region</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>Regional category</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>B&amp;B region</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>Canterbury y/n</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>Longitude</td>
<td>author</td>
</tr>
<tr>
<td>Metadata</td>
<td>Latitude</td>
<td>author</td>
</tr>
<tr>
<td>Data</td>
<td>Q1 Movies</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Data</td>
<td>Q2 Bach</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Data</td>
<td>Q3 Jersey</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Data</td>
<td>Q4 Tag</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Data</td>
<td>Q5 Pause</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Data</td>
<td>Q6 Wagging</td>
<td>survey (author standardized)</td>
</tr>
<tr>
<td>Data</td>
<td>Q7 Loner</td>
<td>survey (author standardized)</td>
</tr>
</tbody>
</table>

3.3 Overview of Collected Raw Data

Between 28 May 2012 and 14 September 2017, 4225 responses to the survey were collected. After establishing that there was negligible variation across the five years within categories, the age brackets were collapsed into a seven-point scale, with the first point containing responses from the youngest participants (born anywhere between 1994 - 2000) and the seventh point containing responses from the oldest participants (born anywhere 1920 - 1945). The numbers in each category also matched onto the real life approximate ages, with '1's
being in their teens, '3’s being in their 30s and so on. The breakdown of New Zealand responses across regions can be seen in Table 4. The eldest group (participants aged 70 years old and over), was eliminated due to low numbers, n = 37, of which 14 were New Zealanders). Figure 9 shows the raw counts by region in a visual format.

Figure 9: Visualisation of raw counts of participants

Table 4: Breakdown of New Zealand responses by age and region

<table>
<thead>
<tr>
<th>Region</th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>70+</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>00-NonNZ</td>
<td>193</td>
<td>525</td>
<td>133</td>
<td>149</td>
<td>92</td>
<td>62</td>
<td>13</td>
<td>1167</td>
</tr>
<tr>
<td>01-Northland</td>
<td>2</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>02-Auckland</td>
<td>69</td>
<td>101</td>
<td>34</td>
<td>33</td>
<td>18</td>
<td>7</td>
<td>0</td>
<td>262</td>
</tr>
<tr>
<td>03-Waikato</td>
<td>21</td>
<td>58</td>
<td>19</td>
<td>10</td>
<td>11</td>
<td>1</td>
<td>1</td>
<td>121</td>
</tr>
<tr>
<td>04-Bay of Plenty</td>
<td>16</td>
<td>28</td>
<td>9</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>66</td>
</tr>
<tr>
<td>05-Gisborne</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>06-HawkesBay</td>
<td>21</td>
<td>26</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>07-Taranaki</td>
<td>14</td>
<td>16</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>37</td>
</tr>
<tr>
<td>08-Manawatu-Wanganui</td>
<td>17</td>
<td>45</td>
<td>13</td>
<td>15</td>
<td>7</td>
<td>6</td>
<td>4</td>
<td>107</td>
</tr>
<tr>
<td>09-Wellington</td>
<td>77</td>
<td>114</td>
<td>42</td>
<td>35</td>
<td>16</td>
<td>6</td>
<td>0</td>
<td>290</td>
</tr>
</tbody>
</table>
From Table 4, we see raw counts of responses, clearly demonstrating the overwhelming majority of respondents born in the Canterbury region aged 19-30. A pleasant way of visualizing data is by using heatmaps, seen in Figures 10 and 11. The numbers reading up the right hand side represent the age categories from people aged 16 - 18 all the way through to people aged 60 - 69. Regions are ordered from North Island to South Island and then North to South within the islands. In Figure 10 we see that 23% of all New Zealand-based participants were born in the Canterbury region aged 19 - 30. We can also see that 35% were born in Canterbury aged 16 - 30. The next biggest representations came from the Auckland and Wellington regions, with the participants from this region aged between 16 - 30 representing seven percent of the total proportions each.

Figure 11 shows a breakdown of region representations within each individual age group. As could be expected, Canterbury represents approximately half of all the responses for people aged 16 - 18 and 19 - 30. It is still the most represented region for all age groups, with the lowest proportion of Cantabrians in the oldest age bracket. Both Auckland and Wellington regions represented around ten percent of all responses across all the age brackets.

As the survey was distributed via the social networks and friends of first year undergraduate students living in Christchurch, New Zealand, the majority of the respondents were aged 19 - 30 (1731 out of 3846). Of those, 1206 were born in New Zealand. There were 2679 responses from New Zealand born participants who provided a region within New Zealand with which we could perform geographical analyses. 0 people responded from the Canterbury region, 32.4% of the total number.

3.4 Cleaning Data

3.4.1 Raw Data Decisions

Of the total number, 3846 responses were analysed. Of the deleted responses, some entries were identical entries by the same participant (possibly from pressing the submit button. The box for entering ethnicity was left blank. Responses that I have marked as ‘New Zealand’ include: kiwi, Pākehā, New Zealand European, Māori, New Zealander, etc. In order for a response to be used, a New Zealand birthplace (which could be sorted into a region) had to be provided, as well as one of the ethnicity terms above.
multiple times), some answers were inappropriate and others were cases where the question had been misinterpreted by the participant. For example, some participants answered using a word in their native (non-English speaking) language, rather than a term in English. Other participants answered the questions with *yes* or *no* instead of providing a term. Where answers were not applicable, they were marked as “NA” and deleted at the analytical stage. Where there were fewer than ten identical terms provided, responses were marked as ‘other’, unless they were directly relevant to Bauer and Bauer’s study (please refer to *Chapter 2: Section 2.3* and *Chapter 4: Section 5*; this is particularly relevant for truce terms). As with all qualitative-turned-quantitative analyses, there was some subjective decisions that had to be made. I outline them below:

- Where an item could be lemmatised (i.e. *wag, wagged, wagging*) all instances of the lemmatised item were treated as identical
- Terms that contained more than one word were condensed to one word using camel case, in order to facilitate analysis with RStudio. (i.e. *playing hooky, picture theatre, holiday home* became *playingHooky, pictureTheatre, holidayHome*)
- Where some words occurred in a number of forms that were not lemmas or different tense iterations, I left them separated (i.e. *time, out* both feature in *timeout*, but are unique entities)
- In two cases I allowed synonyms to be collapsed for ease of analysis. *noFriends* and *noMates* both became *noMates*, the larger of the two. Additionally *holidayHouse* and
holidayHome became holidayHome. However, the terms holidayHouse and vacationHouse were kept separate as the primary words could be regional.

- Where three or more terms were given for any given question, the response was treated as ‘listing’ and was removed from the analysis.

There was naturally a lot of orthographic variation. Homophonous items with unambiguously identical meanings were standardized in spelling. For example; cardy/cardie both became cardigan and hoody/hoodie both became hoodie. The question of orthography was a real challenge for some participants who had perhaps never seen the term written down. Many commented in the boxes with statements relating to this. For example, for the question regarding holiday homes, these statements were provided by three participants “bach (batch? I’ve never known how to spell it!)”, “batch (does it have a ‘t’?)”, “batch (possibly ‘bach’??)”, showing the ambiguity in orthographic convention. The most difficult question to standardize regarding spelling was the question concerning truce terms, for two reasons. The first was the lexical item that could be pronounced ‘pax, pags, bags or pegs’. I kept these separate however some participants kept them together, for example: “pags (bags?)”.

### 3.4.2 Data Management

Once the data had been standardized, it needed to be organized into a format that could easily be analysed using R (R Core Team (2017)). Some people provided up to three unique
entries per question, which needed to be split. I used an R package called tidyverse (Wickham (2017)) in line with the examples provided by Wickham & Grolemund (2017), which allowed me to split the answer column into as many sections as required (in this case three), and give each entry its own row. This meant, for example, that Participant A could have three entries for the question regarding clothing, but Participant B might only have one. The following tables (Tables 5,6, and 7) are examples, not drawn from actual data.

This is an example of data in standardized but not yet tidied form.

Table 5: “Messy” Data

<table>
<thead>
<tr>
<th>Participant</th>
<th>Question 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>jersey, jumper, sweatshirt</td>
</tr>
<tr>
<td>B</td>
<td>jersey</td>
</tr>
<tr>
<td>C</td>
<td>cardigan, hoodie</td>
</tr>
</tbody>
</table>

In this table we can see the participant ID and their response to a particular question (Question 3). However, the responses are contained within a single cell, meaning that each participant will be coded as having different responses, when in fact both participants A and B said they used the term jersey.

Table 6: Converting the data

<table>
<thead>
<tr>
<th>Participant</th>
<th>Question 3ans1</th>
<th>Question 3ans2</th>
<th>Question 3ans3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>jersey</td>
<td>jumper</td>
<td>sweatshirt</td>
</tr>
<tr>
<td>B</td>
<td>jersey</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>C</td>
<td>cardigan</td>
<td>hoodie</td>
<td>NA</td>
</tr>
</tbody>
</table>

In this table the unique words have been separated into three columns. However, it is still difficult to access the individual lexical items as there are now up to three cells per row for the same question. What is more, three of the nine cells here contain no data, which is creating a larger file with unnecessary space.

Table 7: “Tidy” Data

<table>
<thead>
<tr>
<th>Participant</th>
<th>Question 3 answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>jersey ans1</td>
</tr>
<tr>
<td>A</td>
<td>jumper ans2</td>
</tr>
<tr>
<td>A</td>
<td>sweatshirt ans3</td>
</tr>
<tr>
<td>B</td>
<td>jersey ans1</td>
</tr>
<tr>
<td>C</td>
<td>cardigan ans1</td>
</tr>
<tr>
<td>C</td>
<td>hoodie ans2</td>
</tr>
</tbody>
</table>
In this table, every lexical token has its own row, allowing it to be analysed independently from the other responses given by the same participant. All of the participant’s data is duplicated for each unique term they provide. This is now ideal for analysis.

### 3.5 Analysis of Data

The following two sections are added to add some helpful information for the reader regarding interpretation of analysis and interpretations.

#### 3.5.1 Classification Trees

As with Bauer and Bauer (2003), I decided to cluster the data semi-automatically with region and other factors as predictors. To create classification trees I used an R package called *party* designed by Hothorn, Hornik, Strobl, & Zeileis (2010).³

This package creates Conditional Inference Trees with binary branches. It is possible to set the number of breaking points manually, as well as the minimum number of data points within each final node (branch). The first step for this stage of analysis was to set a randomly generated seed (number) which created reproducible results. All of the classification trees used the same syntax, with the replacement of the question-specific Dependent Variable (DV) and the question-specific dataframe. For all other intents and purposes the modelling works similarly to mixed effect regression models, with random effects and a note of significance above 95% probability that the result is not due to chance.

The syntax I used was as follows:

```r
Qnumtree <- ctree(Qnumresponses ~ n_s + age + gender + bbregion + regcat + canterbury + MoveNZyn + MoveOEyn + (1|answers) + (1|Participant), data=dataframe, controls = ctree_control(mincriterion = 0.95, maxdepth = 3, minbucket = 50))
plot(Qnumtree)
```

*Qnumresponses* is the dependent variable, with the independent variables following after the tilda symbol (~). The controls under the *ctree_control* section of the model were the same for all the questions with a maximum depth of three and a minimum number of 50 observations in the terminal leaves of each branch. Random variables were the participant’s ID and the answer column generated from tidying (see Table 7), as items were alphabetized at the standardization stage.

The output, when plotted, looks like Figure 12. The way to read it is to start at the top breaking point (in this case *regional category*) which is labeled Node 1. Following the branches

³Pairwise correlation tests as well as mixed effects regression models were run during the analysis period. However, it was decided that for the final output the classification trees and other plots (see below) were sufficient to convey the results.
past maximum three breaking points, as specified in the model, a bar plot is generated across all the divisions, showing the likelihood of each occurrence. The number of data points in each leaf can be read beside the node label. Where regions are represented as numbers, in this figure, a reminder will be provided in the text. The numbers work low to high, north to south, and can be referred to in full in Table 2, which appears in *Chapter 3: Section 2.1*.

The benefit of using a classification tree as seen in Figure 10 is that multiple significant factors can be seen in a single output. For this study, which had categorical dependent variables, barplots were the simplest visual output to have at the terminal nodes. It is also possible with these classification trees to have other statistical outputs such as a stem-and-leaf plot for numeric results. A downside to these classification trees is that after four lexical items (in this study), it becomes difficult to view the results. The easiest visualization is with two lexical items, so that the results at the end of the ‘leaves’ appear in a stacked bar plot.

![Classification Tree for Question 4 with >100 responses](image)

**Figure 12: Example of a classification tree**

### 3.6. Presentation of Data

#### 3.6.1 ggplot2

The package *ggplot2* designed by Wickham (2009) was used to make the more complex visualizations. Figure 13 shows an example of a bean plot which has several elements. The *y*-axis shows the age category - the higher the number the older the participant. It is fitted to a logarithmic scale and the age was treated as continuous. Each number represents one age
group so the ‘1’ represents the youngest age group, the 16-18 year olds etc. The beans are smaller when there is a smaller concentration of data points for that age/island. The $x$-axis shows the lexical item, in this case the words participants used when replying to Question 6, regarding the word they used, when stating they were missing school illegally. The left and right sides of the plots (shown in blue and red) demonstrate the variation between two factors. In this case and throughout the document the sides represent the North and South islands respectively. We can see a higher concentration on people who *skived* coming from the South Island because the collection of data points is greater on the right hand side. The thick black lines demonstrate where the mean falls for each group. For example, we can see that the mean age for *ditching* in both islands was significantly younger than the mean age for *playing hooky*. This is a good beginning point for examining the data because clear trends for age and island are visible.

![Figure 13: An example of a bean plot](image)

Figure 13: An example of a bean plot
4. Results

This chapter will evaluate the seven questions individually. The discussion regarding the wider implications of the results can be seen in Chapter 5.

Results will be split into the seven questions that were investigated:

Section 4.1. What do you call the public place you go to watch a film?

Section 4.2. Do you have another name for a holiday home?

Section 4.3. What is the name of the item of clothing, often woollen and with long sleeves, that you might wear over a t-shirt to keep warm?

Section 4.4. When you were at school, did you play a game with many players where one player has to run and touch another player while all the other players try to run away and not get touched? If so, what was the game usually called?

Section 4.5. Was there a word which you could say to show that you were not playing for a short time, for instance because you needed to tie up your shoelace?

Section 4.6. If you missed school without permission, for example to go to the park or into town, what word would you use to describe it?

Section 4.7. A person who didn’t have any friends was called what?

4.1 “The place you go to watch a film”

4.1.1 Initial Observations

Unlike other questions, there is a sparse amount of literature regarding the use of this word in New Zealand. Vine (1999:17) noted that the majority of Wanganui participants in her study used the term *movies* although also ‘reported using the British alternative’ (provided later as *cinema*). Bauer & Bauer (2003) did not ask the children this question.

In terms of the raw data, it was clear that *movies* is the most common term for the majority of New Zealanders, accounting for 77% of all responses. This can be seen in Table 8.

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>cinema</td>
<td>112</td>
<td>175</td>
<td>49</td>
<td>31</td>
<td>19</td>
<td>7</td>
<td>393</td>
</tr>
<tr>
<td>flicks</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>11</td>
<td>4</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>movies</td>
<td>500</td>
<td>1020</td>
<td>285</td>
<td>198</td>
<td>103</td>
<td>37</td>
<td>2143</td>
</tr>
<tr>
<td>movieTheatre</td>
<td>26</td>
<td>31</td>
<td>16</td>
<td>7</td>
<td>13</td>
<td>6</td>
<td>99</td>
</tr>
<tr>
<td>other</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>pictures</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>22</td>
<td>16</td>
<td>10</td>
<td>64</td>
</tr>
<tr>
<td>pictureTheatre</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>9</td>
</tr>
</tbody>
</table>

42
For this question, 2780 valid responses were collected from New Zealand born participants. At this stage Other, theatre and pictureTheatre were removed as there were fewer than 50 responses. Pictures was the preferred term for 16% of people in their 60s, but the preferred term for less than 1% of the teenagers. From the beanplot in Figure 14, we see the split between older and younger terms, with flicks, pictures and pictureTheatre all being used by older speakers. The remaining lexical items all were used by participants with an average age of between 19 - 30.

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>theatre</td>
<td>14</td>
<td>16</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>43</td>
</tr>
<tr>
<td>Sum</td>
<td>659</td>
<td>1254</td>
<td>361</td>
<td>275</td>
<td>164</td>
<td>67</td>
<td>2780</td>
</tr>
</tbody>
</table>

Figure 14: Distribution of Q1 terms across age and island

4.1.2 Age and Regional Distributions

The heat map in Figure 15 shows that out of all the New Zealand-born participants, people from Wellington were the least likely to go to the “movies” (70%), and the most likely of all the regions to go to the cinema (25%). Gisborne was the region with the highest proportion of people who went to the movies (92%), although it should be remembered that Gisborne has a very low population count (n=12). People not born in New Zealand were only just
more likely to say *movies*, although 39% of non-New Zealand born participants said they go to the *cinema*.

![Image](image_url)  
**Figure 15:** Proportional distribution of film terms across region and other countries

After viewing the distribution in the beanplot, a classification tree, Figure 16, was created, which showed the distribution of terms that had more than 50 responses. The model for this tree was:

```r
A1tree <- ctree(A1responses ~ n_s + age + gender + bbregion + regcat + canterbury + MoveNZyn + MoveOEyn + (1|Participant), data=A1small, controls = ctree_control(mincriterion = 0.95, maxdepth = 50, minbuckt = 50))
```

Of the dependent variables, only age was significant, with the first significant split happening between the youngest three age groups (participants aged between 16-39) and the three oldest groups (participants aged 40 +). The older participants, whose responses can be seen in node 2, were more likely to go to the *pictures* whereas younger speakers were more likely to go to the *cinema*. Overall though, all age groups had *movies* as their preferred term.

Figure 17 shows the proportions of the top three lexical items for this question: *movies*, *cinema* and *movieTheatre*. From this figure, it is easy to discern that the Wellington region had the highest proportion of the lexical item *cinema*, and Taranaki had the largest proportion of *movie theatre*.
4.1.3 Summary

It seems overall then, that this lexical item is influenced more by age than by region. Although there is one principal term in New Zealand, other names for this action such as going to the *pictures* seems to have faded with time, whilst going to the *cinema* has increased in popularity. The only literature discussing the use of these terms in New Zealand suggested that *movies* was the undisputed favoured term in the country, which is supported by these results. There could be room for debate that some people misinterpreted the question, and named the literal room where one goes to watch a film. Indeed, the largest film-showing companies in New Zealand call themselves Hoyts *Cinema*, Events *Cinema* and Reading *Cinema*. Additionally, when one enters this place, they are told that their film will be showing in *Cinema* N. This exposure could have an effect on the younger respondents, who could most likely go to these places more frequently on average than older respondents.

Looking more closely at the dataset itself in terms of the people born outside of New Zealand, participants born in the United Kingdom were more likely to say *cinema* than *movies*, and people born in the United States were more likely to say *movies* than *cinema*. This aligns with the findings in Vine (1999). Additionally, Australian-born respondents said *movies* approximately 58% of the time, and *cinema* 38% of the time.
Figure 17: Proportional distribution of top Q1 terms across region
4.2 “Another name for a holiday home”

4.2.1 Initial Observations

It is common for families of a higher socio-economic status to either own or hire a second house when they travel around the country for a holiday. Where participants worldwide used labels such as cottage, chalet or lake house, to name a few, this question elicited two other favoured responses from participants: bach (n=2436) and crib (n = 162) (although the terms beach house and holiday home also appeared). A small proportion of people, (n=364) who were born overseas also used these terms, however most of these people had lived in New Zealand for many years. As mentioned in the methodology section, there was some discrepancy over the spelling of the first word, with comments on how people were never too certain about the spelling. Overall, 2633 valid responses from New Zealand born participants were used for this question, approximately 180 fewer than for Question One. It should be noted that if participants responded with individual-related terms, these answers were binned into the ‘other’ category. For example, some people said ‘the house in location’ or ‘family member’s house’. Many people commented on the social status or family wealth needed to be able to use such a place, and therefore term. Table 9 shows a breakdown of responses across the age categories for terms that had more than 10 occurrences and were provided by New Zealand-born participants.

Table 9: Raw Counts of Question 2 responses

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>bach</td>
<td>539</td>
<td>1058</td>
<td>302</td>
<td>236</td>
<td>132</td>
<td>52</td>
<td>2319</td>
</tr>
<tr>
<td>beachHouse</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>28</td>
</tr>
<tr>
<td>crib</td>
<td>13</td>
<td>47</td>
<td>31</td>
<td>14</td>
<td>18</td>
<td>8</td>
<td>131</td>
</tr>
<tr>
<td>holidayHome</td>
<td>43</td>
<td>43</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>91</td>
</tr>
<tr>
<td>other</td>
<td>18</td>
<td>31</td>
<td>2</td>
<td>5</td>
<td>8</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>Sum</td>
<td>622</td>
<td>1187</td>
<td>343</td>
<td>260</td>
<td>161</td>
<td>60</td>
<td>2633</td>
</tr>
</tbody>
</table>

Before beginning analysis, there was some indication that the term crib would be found further south, and the term bach applying everywhere else. Indeed, 140 of the 168 mentions of crib came from Southland region and 19 from the Canterbury region. Figure 18 demonstrates a strong South Island influence for the word crib, especially in the older participants. Bach appeared 2436 times across all ages and across both islands. Holiday Home was a newer lexical term, used by people in their thirties and under. The Southland influence of the word crib seems unsurprising given the evidence provided in Bartlett (2002).

4.2.2 Age and Regional Distributions

The heatmap displayed in Figure 19 shows us several key things. This map was generated for terms that appeared more than 80 times across the entire dataset. The first point of interest
Figure 18: Distribution of Q2 terms across age and island
in this figure is that the most popular term (more than 50%) for the majority of all New Zealanders with the exception of Southland is *bach*. Three regions had this term as their only choice (Gisborne, Marlborough and the West Coast). The second thing to note is that *crib* is almost entirely represented by the two southernmost regions, Otago and Southland. New Zealanders from the North Island were also inclined to use the term *beach house*, perhaps due to the warmer weather overall in the North Island as well as easier accessibility to beaches than in the South Island. Whilst New Zealand born and non New Zealand born respondents sometimes used the term *holiday home*, only those born overseas ever called the building a *vacation house* (24% of the time).

Figure 19: Proportional distribution of second-house terms across region and other countries

The classification tree in Figure 20 shows us that there is a significant split between participants from the southernmost regions (Otago and Southland, see node 9). If people had never moved outside of the city they were born in (see node 10), they were most likely to say *crib* to describe their holiday home than those that had moved (see node 11. If the participant had moved elsewhere in New Zealand they were more likely to behave more similarly to those from other areas in New Zealand, with *bach* as the preferred term. We can see in the tree that if the participant was over 30 and had moved in New Zealand there was a slight inclination to use the word *crib* (see node 7), but zero participants who had not moved used this term (see node 8). It is reasonable to assume that the participants using *crib* may have moved to Southland or Otago at some stage.

The pies in Figure 21 show clearly the weight of *crib* in the southernmost regions, Otago and Southland. We see slivers of this term appear elsewhere in very small proportions. The third
Figure 20: Statistical distribution of the top two second-house terms

most popular term, *Holiday House*, appears in all of the densest populated regions, with small concentrations in the three most southern regions, and the three most central regions.

### 4.2.3 Summary

Unlike the question before it, the choices people made when selecting a lexical item seemed to be slightly more influenced by the region, with people from Otago and Southland (Bauer and Bauer’s “Southern Region”) electing to use the term *crib* where the rest of the country preferred *bach*, for the most part. *Holiday Home* appeared to be a newer term used only by people in their thirties and younger.
Figure 21: Proportional distribution of top Q2 terms across region
4.3 “Item of clothing with long sleeves used to keep warm”

4.3.1 Initial Observations

People were particularly opinionated for this question, insofar as particularly defining a number of terms that seemed to fit the category of ‘a long-sleeved top, often woollen’. This made it difficult to analyse properly, as many people gave answers stating that it depended on whether the clothing item was woollen (despite the description in the question!), whether it buttoned up at the front or whether it had a hood. As with the other questions, responses with fewer than 10 occurrences were automatically moved to ‘other’ in the dataset. Overall there were 2857 responses to this question. Table 10 shows a breakdown of the raw counts.

Table 10: Raw Counts of Question 3 responses

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>cardigan</td>
<td>28</td>
<td>47</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>3</td>
<td>111</td>
</tr>
<tr>
<td>hoodie</td>
<td>15</td>
<td>29</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>51</td>
</tr>
<tr>
<td>jacket</td>
<td>13</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>jersey</td>
<td>350</td>
<td>784</td>
<td>252</td>
<td>185</td>
<td>120</td>
<td>42</td>
<td>1733</td>
</tr>
<tr>
<td>jumper</td>
<td>158</td>
<td>244</td>
<td>72</td>
<td>61</td>
<td>25</td>
<td>8</td>
<td>568</td>
</tr>
<tr>
<td>other</td>
<td>28</td>
<td>25</td>
<td>9</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>sweater</td>
<td>48</td>
<td>76</td>
<td>9</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>144</td>
</tr>
<tr>
<td>sweatshirt</td>
<td>40</td>
<td>71</td>
<td>12</td>
<td>15</td>
<td>7</td>
<td>5</td>
<td>150</td>
</tr>
<tr>
<td>Sum</td>
<td>680</td>
<td>1289</td>
<td>372</td>
<td>284</td>
<td>168</td>
<td>64</td>
<td>2857</td>
</tr>
</tbody>
</table>

For this question there were 2857 responses from New Zealand born participants. The beanplot, as shown in Figure 22 shows all the responses by age. Unlike the first two questions, there does not appear to be a strong influence of age - the mean ages for all terms across both islands hovers between 1.5 and 2.5.

4.3.2 Age and Regional Distributions

The heatmap, as shown in Figure 23 shows the responses given by upward of 100 participants. The most popular terms for New Zealand born participants were jersey followed by jumper, which appeared significantly more frequently than the rest: jersey (n=1842) and jumper (n=594). Northern New Zealand (here Northland and Auckland) were the only two regions that did not have jersey as the favourite term. Gisborne was the only region in New Zealand to have a somewhat strong use of the term sweater accounting for nearly a fifth of responses, although it should be pointed out again that Gisborne had a low population count in this study.

The classification tree in Figure 24 shows a regional split apparent in the heatmap but not the bean plot. Participants born in the two northernmost regions (Northland and Auckland),
Figure 22: Distribution of Q3 terms across age and island

Figure 23: Proportional distribution of warm-top terms across region and other countries
if they had not lived elsewhere in New Zealand, were most likely to use the term *jumper* (see node 5). Additionally, participants from the Waikato and Wellington regions (numbered three and nine respectively) were more likely to use *jumper* than others around the country (see node 6). A note of interest is that participants who were aged between 16-30 at the time of taking the survey were more likely to say *jumper* than their older counterparts, perhaps hinting at a change in progress (see node 11), whilst the oldest speakers, given that they had never moved region, were the least likely to use *jumper* (see node 10).

In Figure 25 we see the distribution of the top two terms, *jersey* and *jumper* across the 15 regions. The majority of the regions had approximately an 80/20 split, however five regions had *jumper* represented more than 25% percent of the time, all in the North Island, and two regions, the northernmost, had *jumper* represented more than 50% of the time.

### 4.3.3 Summary

This question also showed both an age and regional distinction. Although *jersey* was the favoured term around the country for all except the northernmost two regions (Northland and Auckland), it could be that there is a change in progress.

Interestingly, the most popular terms for speakers born outside of New Zealand was *jumper*, the word preferred by people born in the largest and arguably most culturally diverse city in New Zealand, Auckland. This may indicate a change in vocabulary towards *jumper*, or simply demonstrate the cosmopolitan interactions of Aucklanders. Southland had the highest rate of people that use *jersey*. 60 people gave both *jersey* and *jumper*, 41 of those born within New
Figure 25: Proportional distribution of top Q3 terms across region
Zealand. Anecdotally, it appears quite a few New Zealanders do not know which term they use, and many believe they use both.

There were also a few more items proposed for this question that had to be removed. Lexical items such as skivvy and thermal were categorized as ‘not applicable’, as these terms do not refer to the item of clothing specified in the question. For future surveys regarding physical objects, an image may be more helpful to the participants. Vine (1999) found the photographic method to be particularly useful. Few studies have examined the jersey/jumper distinction, instead conflating the two and treating them as a single item, for example Vine (1999) and Barker-Collo et al. (2002). This makes it difficult to draw comparisons.
4.4 “Game where a player has to run and touch other players”

4.4.1 Initial Observations

The question referring to the children’s game was one of the key questions L. Bauer & Bauer (2003) found regional differences for. Several terms were eliminated at the standardization stage, such as *barbadore* and *bullrush* as these terms are used to apply to a different albeit similar game *catch* and *chase* also appeared, but at too low a frequency to be analysed. Overall there were 2670 responses from New Zealand born participants. Table 11 shows the breakdown of raw responses.

Table 11: Raw Counts of Question 4 responses

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>chasey</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>chasing</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>15</td>
<td>15</td>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td>other</td>
<td>37</td>
<td>61</td>
<td>26</td>
<td>47</td>
<td>28</td>
<td>15</td>
<td>214</td>
</tr>
<tr>
<td>tag</td>
<td>566</td>
<td>1050</td>
<td>268</td>
<td>167</td>
<td>78</td>
<td>31</td>
<td>2160</td>
</tr>
<tr>
<td>tig</td>
<td>0</td>
<td>14</td>
<td>10</td>
<td>4</td>
<td>6</td>
<td>3</td>
<td>37</td>
</tr>
<tr>
<td>tiggy</td>
<td>30</td>
<td>78</td>
<td>35</td>
<td>20</td>
<td>18</td>
<td>1</td>
<td>182</td>
</tr>
<tr>
<td>touch</td>
<td>9</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>21</td>
</tr>
<tr>
<td>Sum</td>
<td>642</td>
<td>1209</td>
<td>348</td>
<td>261</td>
<td>154</td>
<td>56</td>
<td>2670</td>
</tr>
</tbody>
</table>

As can be seen in Figure 26, *chasey* (n=20) and *chasing* (n=36) seemed to be older terms, used respectively in the North and South islands. As found in Bauer and Bauer’s study (2003), the three predominant terms for the game for younger people were *tag* (n=2241), *tig* (n=37) and *tiggy* (n=182), with *touch* (n=21) also represented.

4.4.2 Age and Regional Distributions

As can be seen in Figure 27, *tag* was the most popular name around New Zealand and the world for this question. The four northernmost regions (Northland, Auckland, Bay of Plenty and Waikato) all showed an additional preference for the word *tiggy*, reinforcing the results seen in L. Bauer & Bauer (2002a). Otago and Southland both had a small preference for the term *tig*, also in line with Bauer and Bauer’s findings. Aside from that, the West Coast showed small preferences for *chasey* and *touch*, although the numbers were relatively small for both the region and the response.

The overwhelming number of responses for this question were for *tag*, accounting for approximately 80% of the responses. Running a mixed effects regression model demonstrated that region (regcat) came out highly significant and correlated with Bauer & Bauer’s responses. The model showed that *tiggy* is more northern than *tag*, which is central, with *tig* being produced in the more southern regions in New Zealand. Although there were five items, the
Figure 26: Distribution of Q4 terms across age and island
terms seemed to split neatly into two ‘camps’: *ch words* (chasey, chasing) and *t_g words* (tag, tig, tiggy). As seen in L. Bauer & Bauer (2002b) also, it appeared people from the northern regions used *chasey* for the earlier half of the twentieth century before changing to prefer *tiggy*.

The classification tree seen in Figure 28 show us that whilst tag was the most popular name for this game, the four northernmost regions (1- Northland, 2-Auckland, 3-Waikato, 4-Bay of Plenty) also had a significant preference for the name tiggy (see node 7), whilst the southernmost regions (14, Otago and 15, Southland) showed a preference for tig (see node 3). If a participant was born and had never lived anywhere else in New Zealand, the only term that appeared of these three was *tag*. If a participant had moved, they were slightly more likely to use the other two terms, although presumably these participant would have moved or spent some time in the regions that used either *tiggy* or *tig*.

If we look at a model that had over 180 responses (*tig* only had 37), we get a distribution that can be seen in Figure 29. Here, people that had both moved around New Zealand and overseas were more likely to say *tiggy* than those that had never moved, and they were still most likely to say *tiggy* if they came from the four northernmost regions. Again we see that if participants had never moved around in New Zealand, and they were not from the far north, they were highly unlikely to use the term *tiggy*, with South Islanders never using this term if they had never moved (see node 8). This is quite important, especially given that it is the biggest bracket, containing responses from 1131 participants).
Figure 30 demonstrates the distribution of the top three terms across the 15 regions. We can see that *tiggy* appeared in 12/15 regions, with a concentration in the north, as previously noted. *Tig* appears in five regions, although only strongly in the two southernmost regions. Gisborne and the West Coast are the only two regions that contain only *tag*, however they are represented with small populations.

### 4.4.3 Summary

It is clear that there is distinct regional variation here among the lesser-used terms. Bauer & Bauer (2002) explain how *tig* is a common word in the UK used to describe the chasing game, with documented use of the term as far back as 1894. Both the ch- words and the t_g words were in use throughout the 20th century, a claim made by Bauer & Bauer (2002), and supported by the data in the current research. As in their dataset, the ch-words were not used in the southernmost regions (Southland and Otago) at all. In the standardized data it was possible to see that 20% of the people who provided *chasey* also provided *tag* (n=4), and one participant said both *chasey* and *tiggy*. For the central region only 11% (n=2) of participants provided both *chasing* and *tag*. Unlike other questions in this dataset, the results for the most popular two terms *tiggy* and *tag* do not seem to be influenced by age. This could be due to the population density in Auckland region. The only age effect we see is in Southland, with the youngest age groups preferring *tig* far less than their elders. However
these numbers are still relatively small, fewer than 40 people in total used the term *tig*. Out of the New Zealand-born respondents that used *tiggy*, 31 also provided *tag* as an option, compared to four who gave both *tig* and *tag*.
Figure 30: Proportional distribution of top Q4 terms across region
4.5 “Not playing for a short time; temporary immunity”

4.5.1 Initial Observations

Many responses were provided for this question (1766). Bauer and Bauer had fewer regions than the current study, yet found a wide variety in responses. Unlike the other questions investigated in this thesis, this question had several popular terms, rather than one or two clear preferred terms. In total there were 18 unique terms provided for this question from New Zealanders that had a count over ten, as can be seen in Table 12. Of these, six had counts higher than 150: out \( (n=220) \), pause \( (n=365) \), pax \( (n=151) \), pegs \( (n=193) \), safe \( (n=177) \) and timeout \( (n=456) \).

<table>
<thead>
<tr>
<th>Term</th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>bags</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>22</td>
</tr>
<tr>
<td>break</td>
<td>4</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>freeze</td>
<td>7</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>hold</td>
<td>4</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>home</td>
<td>4</td>
<td>10</td>
<td>8</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>notPlaying</td>
<td>8</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>other</td>
<td>16</td>
<td>46</td>
<td>10</td>
<td>6</td>
<td>9</td>
<td>3</td>
<td>90</td>
</tr>
<tr>
<td>out</td>
<td>43</td>
<td>98</td>
<td>29</td>
<td>34</td>
<td>14</td>
<td>2</td>
<td>220</td>
</tr>
<tr>
<td>pags</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>pause</td>
<td>173</td>
<td>176</td>
<td>14</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>365</td>
</tr>
<tr>
<td>pax</td>
<td>1</td>
<td>6</td>
<td>12</td>
<td>48</td>
<td>40</td>
<td>18</td>
<td>125</td>
</tr>
<tr>
<td>pegs</td>
<td>71</td>
<td>96</td>
<td>18</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>193</td>
</tr>
<tr>
<td>quit</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>safe</td>
<td>26</td>
<td>86</td>
<td>35</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>177</td>
</tr>
<tr>
<td>time</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>23</td>
</tr>
<tr>
<td>timeout</td>
<td>105</td>
<td>196</td>
<td>50</td>
<td>30</td>
<td>16</td>
<td>5</td>
<td>402</td>
</tr>
<tr>
<td>truce</td>
<td>5</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>wait</td>
<td>5</td>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>Sum</td>
<td>485</td>
<td>773</td>
<td>194</td>
<td>183</td>
<td>91</td>
<td>32</td>
<td>1758</td>
</tr>
</tbody>
</table>

These can be seen in Figure 31, where a strong age influence can be seen for pax, the Latin term for peace. There does not at this stage appear to be a strong island influence. Out appears to be somewhat favoured for older speakers in the North Island, with timeout being the compliment for older speakers in the South Island.
4.5.2 Age and Regional Distributions

The heatmap in Figure 32 does not show very strong influences of a single preferred term, unlike other results. No term has a concentration level of over 50% in any region. People born outside of New Zealand showed a preference for timeout, the preferred (but not majority) variant in the Auckland, Waikato, the Bay of Plenty, Hawkes Bay, Taranaki, Nelson-Tasman, Canterbury and Otago regions. Pax was the least favoured term for most of the northern regions, as well as Canterbury and Southland. Four regions had pegs as their least favourite term.

In Figure 33, we can see that age is the most influential factor for this term. The lexical items here were the ones that appeared in the New Zealand dataset more than 150 times. It is evident at first glance that pax was the most popular term for the older speakers (aged 40 and over) who were born in the South Island (see node 4). North Islanders in the same age bracket showed stronger preferences for out or timeout (see node 3). For the younger participants, we see in the increase in popularity of pause as the ages decrease, and the use of the term pax gradually decreases to disappear entirely from the responses of the youngest speakers (see nodes 6 - 9).

We see a breakdown of the top three terms (out, pause and timeout) in Figure 34. Unlike other questions, clear patterns are not apparent. Three regions have approximately 50% or more occurrences of pause. Only Northland has approximately 50% out, whilst Waikato and Otago have close to 50% occurrences of timeout. Gisborne is the only region not to have an occurrence of timeout.
Figure 32: Proportional distribution of immunity terms across region and other countries

Figure 33: Statistical distribution of the top six immunity terms
Figure 34: Proportional distribution of top Q5 terms across region
4.5.3 Summary

This question had the most distributed responses across region and age, with no distinct favoured variant. Age was a very influential factor, especially when exploring the six top options. The term *pause* has increased over time, from zero representations in the oldest two age groups to 36% in the youngest age group. Although the majority of the responses were not discussed in Bauer & Bauer’s findings, there were a few minor correlations. Five people in this study used the word *fans*, all of them coming from Wellington across a number of ages, which, although a small number, corresponds with the Playground Talk study. Additionally, *nibs* occurred ten times, eight from the Southland/Otago regions. From Aucklanders, *gates* (n=5) and *quitsies* (n=4) both appeared, although these terms were used once by individuals from the Bay of Plenty, Waikato, Hawkes Bay, Canterbury and Wellington regions as well. Bauer and Bauer found that people from the Nelson/Marlborough area used the term *taxes*; two people from those regions used *tax* and *tags*, which are phonetically similar. *Barley* occurred 34 times in the overall dataset, but not one participant that used this term came from New Zealand.
4.6 “Missing school without permission”

4.6.1 Initial Observations

This question referred to illegally missing a commitment - here missing school. This question was perhaps misinterpreted by some participants who provided the terms *absent* and *truant* in response to the questions. These lexical items were deleted on two grounds: lexical category and then semantic meaning. The majority of terms were verbs, i.e. you *x* school or in some cases you *x from* school. Absent is an adjective. Truant is a noun used to describe the person who is *x*ing school. Table 13 shows a distribution of the raw counts, with a total number of 2715 responses. We can see that terms such as *ditching* and *skipping* school were only used by the younger respondents. Two responses accounted for 90% of all responses given: *bunking* and *wagging*.

Table 13: Raw Counts of Question 6 responses

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>bunking</td>
<td>329</td>
<td>644</td>
<td>137</td>
<td>96</td>
<td>56</td>
<td>16</td>
<td>1278</td>
</tr>
<tr>
<td>ditching</td>
<td>44</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>other</td>
<td>25</td>
<td>24</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td>playingHooky</td>
<td>2</td>
<td>10</td>
<td>7</td>
<td>10</td>
<td>10</td>
<td>7</td>
<td>46</td>
</tr>
<tr>
<td>skipping</td>
<td>32</td>
<td>40</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>81</td>
</tr>
<tr>
<td>skiving</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>wagging</td>
<td>223</td>
<td>485</td>
<td>191</td>
<td>148</td>
<td>87</td>
<td>36</td>
<td>1170</td>
</tr>
<tr>
<td>Sum</td>
<td>656</td>
<td>1236</td>
<td>344</td>
<td>262</td>
<td>157</td>
<td>60</td>
<td>2715</td>
</tr>
</tbody>
</table>

Figure 35 visualizes the age trend seen in Table 13. The oldest terms for this questions were *playing Hooky* and *skiving*. The youngest term was *ditching*. There are two dominant island trends here, seen in a higher representations of *playing Hooky* in the North Island, and higher representation of *skiving* in the South Island.

4.6.2 Age and Region Distributions

In Figure 36 we see the two most popular terms in New Zealand are *bunking* and *wagging*. At first glance, it appears that *wagging* is the most popular of the two, with concentrations on the the right hand side of the heatmap. Taranaki showed a slight predisposition to *ditching*, unlike any of the the other regions. Canterbury had the strongest proportion of bunkers with 81%.

The classification tree seen in Figure 37 shows a model containing the top three responses. It demonstrates first of all a significance of whether or not a participant was born in the Canterbury region or not. If they had only ever lived in Canterbury, they were likely to use *bunking* approximately 90% of the time (see node 4). If they had moved, they were more
Figure 35: Distribution of Q6 terms across age and island

Figure 36: Proportional distribution of truancy terms across region and other countries
likely to use *wagging* approximately 40% of the time (see node 3). This would make sense, as elsewhere in the country with the exception of Taranaki, the most popular term was *wagging*. Non-Cantabrians who had lived elsewhere in the country other than their birthplaces were more likely to use *bunking* than those that had never moved (see node 9). The youngest age groups, irrespective to whether or not they had moved, were significantly more likely to use *skip* school than the older generations (see nodes 8 and 10).

Figure 37: Statistical distribution of top three truancy terms

Figure 38 shows the statistical distribution just for the top two terms. The first change was that Taranaki was included in the first split. Respondents from Canterbury and Taranaki were the most likely to *bunk*, especially if they had never moved in New Zealand. For participants that had never moved and were not born in either of those two regions, nor Northland (1), Gisborne (5) or Otago (14), the likelihood of using the term *wagging* was over 90%. If respondents outside of Canterbury and Taranaki had moved around New Zealand and were in their twenties or younger they were more likely to say *bunking* than their older counterparts. There can be some assumption that a number of these respondents may have moved to Canterbury where the majority of participants came from, and where the survey was distributed from.

Figure 39 demonstrates the proportional breakdown of *bunking* and *wagging* across regions. Two regions show a preference for *bunking*; Taranaki and Canterbury. We also see somewhat strong preferences coming from the Northland, Gisborne and Otago regions. Participants born in the Hawkes Bay region were the least likely to *bunk*.
4.6.3 Summary

This question did not appear in Bauer and Bauer’s study, despite it being a childhood term. It appears that this term is influenced both by region and age, with *bunking* used more among the younger speakers with a concentration in Canterbury. Newer terms also seem to be on the rise with *ditching* and *skipping*, with *playing hooky* a term of the past.
Figure 39: Proportional distribution of top Q6 terms across region
4.7 “Someone with no friends”

4.7.1 Initial Observations

This question appeared in L. Bauer & Bauer (2003)’s study for a child that appeared to have no friends. A variety of answers provided had to be deleted as they were too specific for the question. For example, 17 people sadly responded *me* in responses to this question. Two used *OTL*, short for *only the lonely* while others commented that there was no word for that in their school (or they were unaware of a term). Yet others listed proper names. The *name*-no-mates was common term that came with a variety of names preceding the ‘no-mates’: *Billy, Johnny, Nancy, Nellie, Noddy, Nobby, Norma, Norman, Scott* and *Nigel* although only *Nigel* appeared more than 10 times. There was also a lot of condensing that had to be made surrounding the term *nf/nif(f)* which people commented could stand for:

- No friends
- No identifiable friends
- No intelligent friends
- No interesting friends

This term could be used as a verb as in ‘*you’ve been niff-ed*’. Bauer and Bauer claimed that this was a Canterbury specific term, which will be tested over time and region. Bauer and Bauer also mentioned that ‘OTL’ or *Only The Lonely* was a term used by children, however this only occurred twice in my data and thus was reduced to the *other* category. Other terms that were moved to the *other* category in the New Zealand only dataset for having fewer than 10 responses included: *geek, outcast, outsider, retard* and *sad*. Many of these terms had higher counts in the overall set, perhaps demonstrating and international, rather than regional, influence. Overall there was a number of 2573 that of responses that were used for the final analysis. As with the previous questions, the raw counts can be seen in the table below (Table 14).

<table>
<thead>
<tr>
<th></th>
<th>16-18</th>
<th>19-30</th>
<th>31-39</th>
<th>40-49</th>
<th>50-59</th>
<th>60-69</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>lonely</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>loner</td>
<td>534</td>
<td>1013</td>
<td>134</td>
<td>133</td>
<td>73</td>
<td>29</td>
<td>1916</td>
</tr>
<tr>
<td>loser</td>
<td>36</td>
<td>86</td>
<td>39</td>
<td>25</td>
<td>21</td>
<td>3</td>
<td>210</td>
</tr>
<tr>
<td>loser</td>
<td>36</td>
<td>86</td>
<td>39</td>
<td>25</td>
<td>21</td>
<td>3</td>
<td>210</td>
</tr>
<tr>
<td>nif</td>
<td>0</td>
<td>31</td>
<td>66</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>nif</td>
<td>0</td>
<td>31</td>
<td>66</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>100</td>
</tr>
<tr>
<td>nigelNoMates</td>
<td>3</td>
<td>20</td>
<td>37</td>
<td>27</td>
<td>9</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>nigelNoMates</td>
<td>3</td>
<td>20</td>
<td>37</td>
<td>27</td>
<td>9</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>noMates</td>
<td>0</td>
<td>6</td>
<td>25</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>noMates</td>
<td>0</td>
<td>6</td>
<td>25</td>
<td>7</td>
<td>3</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td>other</td>
<td>30</td>
<td>34</td>
<td>14</td>
<td>24</td>
<td>14</td>
<td>6</td>
<td>122</td>
</tr>
<tr>
<td>other</td>
<td>30</td>
<td>34</td>
<td>14</td>
<td>24</td>
<td>14</td>
<td>6</td>
<td>122</td>
</tr>
<tr>
<td>reject</td>
<td>2</td>
<td>3</td>
<td>15</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>weirdo</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>weirdo</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>16</td>
</tr>
<tr>
<td>Sum</td>
<td>622</td>
<td>1208</td>
<td>336</td>
<td>231</td>
<td>131</td>
<td>45</td>
<td>2573</td>
</tr>
</tbody>
</table>

4 The phrases *nigel no-mates* and *no-mates Nigel* were collapsed into one.
Table 14 show us that the most popular term by a long way is *loner* (n=1916) followed by *loser* (n=210). Some people made the distinction of the split being by choice; 31 participants gave both terms as options. The common description was that a *loser* was more unlikable and had no friends through personality, whereas a *loner* was someone who chose to be alone rather than with friends.

Figure 40 shows the distribution of the responses across age and island. As with Question 5, this question had multiple response given, however this time there was one preferred term and three other terms that had counts of between above 50. Older terms with higher age means were *Nigel no-mates* in the North Island and *reject* in the South Island. The youngest terms were *weirdo* and *lonely* in the South Island and *loner* in the North Island.

**4.7.2 Age and Region Distributions**

We can see the most popular terms easily in Figure 41. As with Question 1, there was one response (in this case *loner*) which accounted for over 70% of responses in all regions, including overseas. Only a handful of the southern regions (West Coast, Canterbury and Southland) show a slight indication of using *nif*. All of the regions with the exception of three (Bay of Plenty, Gisborne, West Coast) also had approximately a 10% representation of *loser.*
As can be seen in Figure 42, age is once again the most significant factor in predicting a lexical choice. With the exception of Cantabrian-born participants (aged between 31 and 39 at the time of the survey, see node 2), the most popular term across the country was *loner*. It appeared only participants aged between 31 - 39 used the term *nif*, perhaps a short fleeting phase. Incidentally, there would have been participants in this age bracket who could potentially have provided answers for Bauer and Bauer’s Playground Talk study (2003), as some would have been aged 11 or 12 in the late 1990’s. One term that appears to have almost disappeared for the younger generations (teenagers and people in their twenties) is the term *Nigel No Mates*. This seemed to be popular for the 31 - 59 year olds, but not for people aged 60 - 69.

In Figure 43 we see the breakdown of the top four terms across region. We can see the presence of *Nigel No Mates* in most of the regions, represented in green. There is a concentration of *nif* in the southern regions, as noted in the heatmap above (see Figure 41). The Otago region has the smallest use of the term *loner* and incidentally the largest use of *loser*.

### 4.7.3 Summary

For this question, age was once again the most influential factor, with *loner* growing in popularity as rival terms such as *Nigel No Mates* and *loser* waned. As discovered in L. Bauer & Bauer (2003), *nif* seemed to be Canterbury based and also remarkably short lived, as demonstrated in Figure 42. It appears that, like the question above, there are relatively
strong influences of both age and region.

Figure 42: Statistical distribution of no-friend responses
Figure 43: Proportional distribution of top Q7 terms across region
5. Discussion

While it is clear that there are majority lexical items that speakers of New Zealand English use, there is also ample evidence for regional variation occurring/emerging. Social factors were also found to be significant. This section will address the three research questions proposed in Chapter 2: Section 4. Section 5.1 in this chapter will evaluate the differences of responses between speakers across the two major islands. Regarding variation across the 15 regions outlined in the Methodology section, Section 5.2 will compare and contrast previous studies to the results identified in this study. Finally, in Section 5.3 I will evaluate the influence of social factors such as age and gender played in the variation that occurred in this dataset.

On a broad level, this thesis adds to the evidence that New Zealand English is still forming, or at the very least evolving, in line with Trudgill et al. (2000). Despite historical research that claims there is no regional variation in New Zealand when the country and its people were discovering their identity, the dialect and nation is now at stage where diversification is now ready to occur. Previous studies have indicated the emergence of particular regional variants (Bartlett, 1992; Quinn, 1995; Starks, 2000; Bauer and Bauer 2003; Schreier et al., 2003, Ainsworth, 2004; Kennedy, 2005; Marsden, 2013). This thesis reinforces the loose division of New Zealand into three linguistic areas, with the newest one, a northern New Zealand division still perhaps in its infancy.

5.1 Island Variation

As will be discussed below in Section 5.2, it would seem that regional variation when it occurs is not influenced by the island that a speaker comes from. This could be considered rather surprising (many other studies from other countries have demonstrated clear splits between the north and southern ends of their country) and given the physical water barrier, until we consider the geographical layout of New Zealand. Across land transport can be quite difficult in New Zealand, and it is easier to get between the Marlborough and Wellington regions than it is to get between Canterbury and the West Coast, especially during winter. Additionally flights between the major cities of New Zealand (Auckland, Wellington, Christchurch and Dunedin) occur several times every day and are significantly less expensive than inter-region flights (for example between Hamilton and Tauranga, which are situated in neighbouring regions. ) It appears that the regional variation occurring across the islands is driven more from the northern and southern areas more so than the islands themselves.

One of the questions in this study that revealed significant island variation was Question 4 (see Chapter 4, Section 4) when evaluating the top two terms. For this question, although tag was the dominant lexical item, the term tiggy was never used by someone who had lived in the South Island their whole life and had never lived anywhere else in the country. Having said this, the term tiggy was more significant in terms of the northern regions of the North Island; participants who had lived south of Gisborne and the Hawkes Bay areas were also

---

5On searching this exact flight, it was discovered that one needs to fly to Auckland first, there are zero direct flights.
highly unlikely to use *tiggy* if they had never lived elsewhere. Another question with an island distinction was Question 5 (see Chapter 4, Section 5), with South Islanders over 40 years old more likely to use *pax* as a truce term and North Islanders over 40 years old were more likely to use *out, safe and timeout*.

### 5.2 Regional Variation

There is enough evidence in this study as well as in previous studies to argue for the existence of regional variation in New Zealand English, albeit in small pockets. This study contributes to the argument for identifiable and statistically significant differences in linguistic behaviour for individuals from different regions. Whether examining lexical variation (current study, Bauer and Bauer), syntactic variation as researched by Quinn (1995) or phonetic variation as documented by Kennedy (2006), Ainsworth (2004), Marsden (2013) and Bartlett (2002) the differences in performance across regions are clear. Whether or not this is purely geographical or whether other factors such as urban vs ruralness or ethnic proportions have a stronger pull has been a contentious rival in this debate.

The strongest regions that consistently appeared as behaving differently to other New Zealand regions were Southland (sometimes behaving the same as Otago) and Northland and Auckland. This supports foremost Bauer & Bauer’s idea of the division of the country into three main linguistic regions, although the actual boundaries may be up for discussion. The most evident example comes from Question 4, where the division into three distinct regional areas is clear. Participants who were born in one of the regions with the main cities (Auckland, Wellington, Canterbury and Otago) tended to give more responses, and of course there were more data points than for the more rural regions.

#### 5.2.1 Northern New Zealand English

Although there is current research investigating an emerging “Auckland” dialect in particular South Auckland with the research of Starks (2000) and Duhamel & Meyerhoff (2015), the present thesis is one of the earliest studies that compares the lexical variation from Auckland and Northland and directly contrasts it with that of lexical decisions made elsewhere in the country. Hopefully the results of the current study can be correlated with the phonetic findings of other research.

#### 5.2.2 Southern New Zealand English

“Southland English” was already a dialect of New Zealand English that was already accepted as existing in lay and linguistic circles. The results support the robust phonetic evidence presented in *Chapter 2, Section 2.3.2* with lexical support. In certain questions, such as Question 2, the Southland difference was very evident. For other questions that were perhaps more influenced by age than region, Southland did not appear so different.
5.2.3 Central New Zealand English

This may not be in itself worthy of being defined as a separate dialect, more the general responses form the majority of New Zealanders. Although certain regions (Canterbury, Taranaki) occasionally behaved differently to their neighbours, the differences were not as consistent as the northern/southern split.

5.3 Social Variation

5.3.1 Gender Variation

Gender was the least influential factor in most of the questions. It would be unwise to suggest that gender never has an effect however. Perhaps this methodology was not conducive to revealing variation, or perhaps gender variation is less significant when it comes to everyday or childhood lexical items. To be fair there, were only seven questions analysed, and perhaps a larger study would reveal gender variation. White and Hay (in prep) discovered that genders had different frequencies for particular lexical items.

It is difficult to compare results for gender differences in lexical choices from other studies. Bauer and Bauer (2002) as well as Kennedy (2006) could not test for gender influences, as their data was collected from group responses.

5.3.2 Change Over Time

This section will briefly summarize the age-related results from the previous discussion before comparing these to other studies and presenting some possible explanations. Age was found to be the biggest social factor influencing lexical variation, as expected. All of the responses showed an interaction between age and lexical item.

Question One showed the disappearance of theatre and movieTheatre with a significant increase in the use of the term cinema. Question Two showed an increase of holidayHouse over time. Among the two most popular options, crib was an older word. Question Three showed a difference that was significant with gender and age in isolation (jersey was used by older people, particularly females, and jumper was used by younger people, particularly males). When this appeared in a classification tree however the age factor disappeared in favour of age, region and whether or not the participant had moved either overseas or within New Zealand. Investigating classification trees again, Question Four demonstrated an age difference for respondents living in the two southernmost regions (Otago and Southland) where speakers aged 30 and over were more likely to use tig than their counterparts. It is evident here that in this case regional variation is not emerging, rather fading. This correlates with L. Bauer & Bauer (2002a), who found that the use of tig was found as high as South Canterbury for older speakers. Tag is the most common term in the USA, and used to be a common term in the UK, which was replaced by tig. As discovered in L. Bauer & Bauer (2002a), chasey and chasing were both regional but crucially age graded, with only the oldest
speakers employing these terms. The results in this study correlate with Bauer and Bauer (2002), highlighting similar geographical and time-divisible variations. Although there were not many instances of touch, this was also a distinctively younger term.

Moving onto the next question regarding truce terms, the oldest term was clearly pax, with very few of the speakers under 29 selecting this lexical item. When only examining the most frequent terms, age was the most important factor. L. Bauer & Bauer (2002a) point out that for their study very few people born before 1930 (the “7”s in this study) had a truce term. Although not included in the statistical analysis for paucity of numbers, this was not the case for the oldest speakers in this study. Of the 14 New Zealand born respondents, four said pax, two said nibs, and there was one observation each of base, peace, out and timeout. This means that four speakers (29%) did not have or provide a truce term. Having said that, 180/638 speakers (28%) aged 16 - 18 also did not have or provide a truce term, so Bauer and Bauer’s claim that truce terms did not exist earlier than the 19twenties are not fully supported by these results. Bauer and Bauer do not mention the two most popular terms in this study: pause and timeout. It was clear to see a decrease over time of the term timeout and an increase of the use of pause. It was only for the youngest speakers that pause was the most favoured variant, but it outnumbered the others by 15% in the youngest category. When investigating the top six most frequent terms, age was once again the leading influential factor, especially for participants aged 39 and under. We note an increase of the term pegs coincidentally with the decrease of pax. Both respondents and Bauer and Bauer (2003) described pegs as a mutation of pax. With the New Zealand short front vowel shift, this makes sense, as the DRESS vowel raised in the vowel space to be phonetically realized as /e/ Gordon et al. (2004). This vowel change only accounts for one of the sound changes (the vowel), creating the word one could write as pex. However, the change did not appear to change in this way. No individual used the word pex, whereas twelve people gave the term pags and 26 gave the term bags so perhaps it was the /ks/ -> /gs/ sound that changed first. As noted by Papp (p.c.) there are not a lot of phonetic differences between these consonants in word-final position, as codas are seldom voiced in New Zealand English. The leap from both pags and bags to pegs could be attributed to two reasons. The first is that pags is not an existing word, whereas pegs is a recognizable object. Children could be exposed to the word pegs so often that they misheard pags or believed what they were hearing was pegs, especially considering the vowel change. Latin has not been taught in the New Zealand Curriculum for many years, the Latin translation of pax to peace may well have been lost. Secondly, as mentioned in L. Bauer & Bauer (2002a) the word pegs can also be used to lay claim to something (as with bags) and a peg is a grippable object that can be used to attach to an object, making it identifiable (and showing a claim laid). Similarly, pegs could be used in a metaphorical sense to lay claim to something. There was an obvious age influence for Question Six with the decrease and disappearance of playing Hooky with an increase of bunking over time (although this was more regional related). Two new terms appeared for the younger speakers also: ditching and skipping. For Question Seven, age was once again the most significant factor, with the youngest respondents almost exclusively employing the term loner, whilst the similar word loser was used more by older speakers.
6. Limitations and future research

6.1 Time

This is a small sample of the possible questions I could have analysed for this thesis. Therefore the first limitation was, of course, time. It would be very useful to have been able to analyse all of the questions in the survey, providing more results with which to compare to Bauer and Bauer (2003)’s study, but also perhaps shed more insight into non-playground related words. This would have provided another broad overview and the current study is able to probe a little deeper. It was not possible before beginning analysis to know which questions were going to be helpful and which were not, although L. Bauer & Bauer (2003) provided a start in the right direction.

An additional difficulty with time was the process of learning all of the necessary and helpful tools needed to undertake this project. Due to the fact that each individual location was not geocoded (again, due to time), data points were only plotted onto 15 points on the map. A visualization method called geom_jitter was used to move the dots slightly but even with this results were difficult to interpret for the reader. A development of this idea was to plot pie charts of the densest regions beside their locations on the map. However, this also became confusing as the proportions of each lexical item were repeated on the map for the larger regions. Time was spent learning to create and then generate maps, violin plots, polar-coordinate graphs and various bar graphs before concluding that pie graphs were perhaps the best was to visualize some of the data, especially since they were a secondary supplement to the main results, in order to help the reader get an overview of the distributions.

As L. Bauer & Bauer (2002b) also noted, if regional variation can be found in the lexical choices of participants, we would also expect to find phonological differences between adults across the three proposed “dialects” of New Zealand English. Small advances have been made, with Bartlett (2002)’s dissertation on the Southland dialect and emerging research regarding an Auckland/Northern New Zealand manner of realising /s/ and /r/.

6.2 Sampling area

This study was distributed to university students and their networks. It would be possible that a survey distributed more equally between tertiary and non tertiary educated participants would reveal different results. Having said that, it would be difficult to distribute a survey on the scale of this project without using at least initially, some of the network of the researchers in order to begin the process. With zero compensation for their time, participants need to be encouraged to fill in the survey for the purpose of helping research, and people are more likely to aid if they know the person undertaking the research.

Alternatively, due to the significantly higher proportions of Cantabrians responding to the survey, it would have been useful to sample the data as in Hay & Foulkes (2016). To mitigate that issue in this project, the majority of results were drawn upon within-region proportions of lexical items. Future research would want to try as much as possible to have equal numbers
from each region and age category, although naturally this is task more easily suggested than achieved.

### 6.3 Open Boxes

Using a dataset created by someone else will always come with some unavoidable difficulties. This paragraph will briefly cover some potential changes for the survey before moving on to the limitations of the study I had more control over. Were I to recreate this survey from scratch it would be perhaps easier to have categorical instead of open boxes for ethnicity and region. This would allow ethnicity to be a variable. For the city it would be useful again to have a box to search your own city, this way it may be possible to geocode the data and examine the relationship between the behaviour of urban and rural dwellers. One more question that would be useful would be “Which city do you live in now?” to allow for the possibility that speakers may behave more like their current city than their birth city.

Even taking into consideration the pre-selection of questions, merely transforming the responses into a tidy form took weeks. As mentioned in Section 3.4.1, orthographical variation was an impeding factor in standardizing responses. Particularly in the childhood section, many of the terms were used only in the playground at break times, and children would never have had much cause to write the terms down or see them written. Some participants clearly got bored or had fun with the questions, writing inappropriate answers to questions, or writing anecdotes. Some comments were useful and made it into the final discussion, however much of it needed to be cut from around the target word. A second issue with this methodology which was admittedly difficult to avoid was the disproportionate number of young (19 - 30 year old) Cantabrian participants that answered this survey. The reason this happened was due to the sociolinguistic group that distributed the survey each year. This issue was mitigated by sampling both the regions and age groups and in the end was not a major problem to deal with. In fact, it provided some insights to the uniqueness of what could arguably be an emerging Canterbury dialect. A variety of unique responses and commentaries were able to be collected and analysed due to the blank boxes. Perhaps in the data collection ask people to add one term, then select to add another term (discourages multiple answer) and allow comments to be added later. This makes standardizing easier but removes some of the comments that allow us to see if the selection is the word the participant themself uses. Although an open box yields diverse responses and allows participants to express the ethnicity they identify with freely, it was not ideal for questionnaire of this magnitude. Indeed, an entire separate thesis could be written on people’s self-labeled ethnicity across regions. For example, Cantabrians were more likely to identify themselves as Europeans, whereas Aucklanders had a higher percentage of self describing as āPākehā. It was also difficult to code for an urban/rural split as evaluated in Marsden (2013) and Duhamel & Meyerhoff (2015), due to the open-box nature of the survey. Some later analysis was performed by splitting the three densest regions into “urban” (Auckland, Wellington, Canterbury) and the rest into “rural” however this was not satisfactory as no questions came up significant. It is presumed to be an issue of method rather than lack of results.
6.4 Future Research

This methodology allowed for the collection of thousands of responses, and also collected important and testable demographic data such as gender and ethnicity (two factors missing from Bauer & Bauer’s study). Due to the focus on lexical, rather than phonetic, data, it was possible to collect data from all over New Zealand and the world, without having to travel to each participant. People could respond to the survey from the comfort of their own homes, resulting in people from all ages and backgrounds participating. The project did not cost money, the participants kindly donated their time. Unfortunately the data is not so comparable that I would reject the divisions provided in Bauer & Bauer (2000) and propose my own. I did not measure responses from individual sub-regional areas. However I did collect social data such as age and gender in order to test the significance of the regional results. On the other hand, Bauer & Bauer were able to put ‘individual’ (this may not be the correct term, as each single response was actually the response of a classroom of up to 30 individuals) data points onto separate areas of a map, and delineate their regions that way. I began with pre-set regions. Whether or not this was the best way to approach the problem is certainly a good point to raise. Given more time to analyse the data, and more knowledge in geo-coding, it would be very useful to plot the data points individually and then cluster them into regions statistically.

To expand the results in this study, it the first starting points would be to analyse all the questions in the survey, or create a different survey that allowed for easier geocoding and standardizing and distribute it to a wider range of respondents (although there is a sizable number already in the current dataset). It would be useful to test for an urban/rural distinction, since there are strong indications that it may be an influential factor. Additionally it would be interesting to see if there is a three-way distinction in spoken New Zealand English, perhaps using a corpus such as ONZE or the Wellington Corpus of Spoken New Zealand English (WCSNZE) which both contain recordings from New Zealanders around the country.
7. Conclusion

There were several key questions asked at the beginning of this thesis. Chapter 2 evaluated previous research regarding regional variation, and posited that a middle ground between qualitative and quantitative research was an appropriate way to address the variation. The current analysis is a more quantitative-leaning take on lexical variation than has been done in the past. Overseas research into lexical variation took into consideration elements such as how regional a person was, or their age. Modern research in New Zealand English has undertaken age, sub-regions and urban/ruralness into consideration, but not gender. Chapter 3 outlined the processes used to transform a raw dataset of survey answers into a tidy analyzable format, how questions were explored and visualized using beanplots, classification trees and maps. The results section of Chapter 4 outlined the most important results for each of the seven survey questions. For many of these questions age and region were significant factors affecting the lexical choice and the details were discussed in Chapter 5. No research completed for a thesis will ever be fully explored due to time, financial and geographical pressures, and it was an overview of some of the limitations and difficulties discovered during this project that appeared in Chapter 6, as well as a suggestion of where future research could lead.

Through the analysis of lexical items used both in childhood and everyday adult situations, I have been able to push the boundaries of knowledge regarding regional variation in New Zealand English out a little further. This study has introduced statistical analysis as a tool for which to explore the correlations between spatial boundaries and social boundaries. It is a project which has implications within geolinguistic and sociolinguistic spheres, and importantly it contributes to what we already know about New Zealand English, reinforcing the idea that lexical variation as well as phonetic variation can be a fruitful area to explore when discussing regional variation, and that the results plotted on a map can be tested to discover which other factors are significantly affecting an individual’s choice. Moreover, the results uncovered during this project supported the findings of past researchers, which put future researchers in good stead of starting points for continuing our work. There is a great amount of ground that has been covered between us all, and yet there is plenty more work on regional variation to be done, especially in light of the idea that regional variation in New Zealand English is still very much in early stages, and will, I believe, continue to develop more fully over the coming years.


Papers in Linguistics, 8, 70–92.


Appendix I: Questions asked in the “Words We Use” survey

Admin 1. Are you male or female?
Admin 2. How old are you?
Admin 3. In which country were you born?
Admin 3a. In which town/city where you born?
Admin 3b. Which suburb are you from?
Admin 4. What is your ethnicity?
Admin 5. If you have lived outside your country of birth for more than three months, where have you lived and for how long?
Admin 6. Did you go to school in the same city/town you were born in?
Admin 7. Which country and city/town did you go to school in if it wasn’t the same?
A1. What do you call the public place you go to to watch a film?
A2. You have bought a mixed bag of things like caramels, wine-gums and toffees. How would you describe these?
A3. You are going to the beach and want to swim in the sea, but you forget the clothes you would usually wear when you do this. What would you call the clothes in question?
A4. Do you have another name for a holiday home?
A5. Do you have a name for the thing that a girl with long hair would wear on their head to stop hair falling onto their face? A7. What is the name of the snack, made from slices of potato, which comes in a bag or packet. You can get them in different flavours (e.g. salt vinegar, chicken, bbq etc)?
A8. What is the name of the item of clothing, often woolen and with long sleeves, that you might wear over a t-shirt to keep warm?
B1. When you were at school, did you play a game with many players where one player has to run and touch another player while all the other players try to run away and not get touched? If so, what was the game usually called?
B2. In the game you identified for question B1, did you have a word to tell someone that they are to be the person who tries to touch the others?
B3. In the game you identified in question B1, was there a special word which the player who touches someone says as they touch them?
B4. Was there a word which you could say to show that you were not playing for a short time, for instance because you needed to tie up your shoelace?
B5. Here is the description of another game: This game needs a lot of players. All the players except one stand behind a line on a field with a parallel line some distance away. One person is in the middle. The person in the middle calls a player by name. That player (the runner) tries to run to the other line without being touched by the player in the middle. At any time the runner can call out a special word, and then all the other players also have to try to run across without being touched. Any person touched stays in the middle and helps the first person. The player in the middle can also make everyone run by calling out a special word. If you played this game in school, what was the game called?
B6. For the game described in question B5, what was the word you used to make everybody run?
B7. Here is a description of another game. One player stands facing a wall. The other players stand behind a line some distance away. They try to creep up and touch the wall. The player facing the wall can turn round at any time. Any player seen moving must return to the base line. The first player to reach the wall takes
the place of the player against the wall, and the game begins again. If you played this game in school, what was it called?

B8. At school, if you said exactly the same thing at exactly the same time as somebody else, did you say a particular word? If so, what was the word?

B9. If you missed school without permission, for example to go to the park or into town, what word would you use to describe it?

B10. At school, someone who is really good at games and sports, but not good at school work might be called a what?

B11. Some people really like school work. Often they also like computers. People like this might be called what?

B12. A popular person or group at school was called what?

B13. A person who didn’t have any friends was called what?

C1. What word do you use to describe somebody who has become intoxicated from too many alcoholic drinks? C2. You have won some money on the national lottery, and you are very happy about this. What word would you use to describe this feeling?

C3. Someone bumps into you at the bus stop and spills your drink all over you. He is very apologetic. You want him to know you are not upset about it. What would you say?

C4. How would you describe something you think is really very good?

C5. How would you describe something you think is really very bad?

C6. It’s -5 degrees outside, and you don’t have a coat. What word would you use to say how cold you are?

C7. How would you describe something you think is very easy?

C8. How would you describe something you think is very difficult?

C9. If something is badly damaged, perhaps beyond repair, how would you say this?

D1. Do you have another word for ‘friend’?

D2. When you see a close friend, what words do you use to greet them?

D3. You and a close friend have spent the afternoon in a cafe, but you are about to leave. What do you say as you leave?

D4. What word would you use to describe an attractive person?

D5. What word would you use to describe an unattractive person?