



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

Research Commons

<http://waikato.researchgateway.ac.nz/>

Research Commons at the University of Waikato

Copyright Statement:

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

The thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of the thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from the thesis.

Tari Rorohiko/Department of Computer Science



THE UNIVERSITY OF
WAIKATO

Te Whare Wānanga o Waikato

Kirikiriāroa/Hamilton, Aotearoa/New Zealand

Indigenous Language Usage
in a Digital Library:
He Hautoa Kia Ora Tonu Ai

A thesis submitted as requirement for the degree of
Doctor of Philosophy at the University of Waikato.

Te Taka Keegan

June 2007



Ka hiki atu taku wae i tēnei whenua
Ka moanatia i muri i a au
Pūrekirereki Wiwi māna tātou
Hei huti ake ki te ao nei
Ka puta ka ora.

Ko Tainui te waka
Ko Pirongia te maunga
Ko Waipā te awa
Ko Waikato-Maniapoto te iwi
Ko Ngāti Apakura,
 ko Ngāti Hikairo ngā hapū
Ko Pūrekireki,
 ko Waipapa ngā marae.

Ko Hikurangi te maunga
Ko Waiapu te awa
Ko Ngāti Porou te iwi
Ko te Whānau-ā-Karuwai
 ki Maraehara te hapū
Ko Te Rehu-ā-Karuwai,
 ko Te Hiku-ā-Mahiti
 ngā tūpuna whare.



Abstract

The research described in this thesis examines indigenous language usage in a digital library environment that has been accessed via the Internet.

By examining discretionary use of the Māori Niupepa and Hawaiian Nūpepa digital libraries this research investigates how indigenous languages were used in these electronic environments in 2005. The results provide encouragement and optimism to people who are striving to retain, revitalise and develop the use of indigenous languages in information technologies. The Transaction Log Analysis (TLA) methods used in this research serve as an example of how web logs can be used to provide significant information about language usage in a bilingual online information system. Combining the TLA with user feedback has provided insights into how and why clients use indigenous languages in their information retrieval activities. These insights in turn, show good practice that is relevant not only to those working with indigenous languages, indigenous peoples or multilingual environments, but to all information technology designers who strive for universal usability.

This thesis begins by describing the importance of using indigenous languages in electronic environments and suggests that digital libraries can provide an environment to support and encourage the use of such languages. TLA is explained in the context of this study and is then used to analyse aspects of te reo Māori usage in the Niupepa digital library environment in 2005. TLA also indicates that te reo Māori was used by international clients and this usage differed to te reo Māori usage by national (Aotearoa) clients. Findings further reveal that the default language setting of the Niupepa digital library had a considerable impact on te reo Māori usage. When the default language was set to

te reo Māori not only were there more requests in te reo Māori but there was also a higher usage of te reo Māori in the information retrieval activities. TLA of the Hawaiian Nūpepa digital library indicated that the Hawaiian language was also used in a digital library. These results confirm that indigenous languages were used in digital library environments.

Feedback from clients suggests reasons why indigenous languages were used in this environment. These reasons include the indigenous language content of the digital library, the indigenous language default language setting of the digital library and a stated desire by the clients to use the indigenous language. The key findings raise some interface design issues and support the claim that digital libraries can provide an environment to support the use of indigenous languages.

Preface

The thesis title, *Indigenous Language Usage in a Digital Library: He Hautoa Kia ora Tonu Ai*, sums up the intent and paramount conclusion of the research undertaken in this thesis. The usage of indigenous languages, specifically te reo Māori and the Hawaiian language, are examined in a digital library context, and the conclusion drawn is that digital libraries can provide an environment of support (he hautoa) for the continued use and subsequent on-going survival (kia ora tonu ai) of these languages.

To investigate an environment which contributes to the retention, revitalisation and development of an indigenous language would be more appropriate if written in that indigenous language. As a speaker and writer of te reo Māoriⁱ I consider this a personal obligation to the inherited treasure that is te reo rangatira (the language of chiefs). The fact that the text of this thesis is written in English and not te reo Māori further marginalises te reo Māori. The decision to write the body of this thesis in English was made reluctantly. Writing the text in English allowed for advice and supervision from experts in the many fields of this research and has made the results more accessible and consequently more relevant on a wider scale. However, I would like to state for the record that when other students or researchers intend to write dissertations about a computer science topic in te reo Māori I will make myself available for assistance and/or supervision whenever possible.

ⁱ Te reo Māori means the Māori language; reo Māori means Māori language.

While the content of this thesis is written in English I have used other languages in the preamble. Following the title page, and before the Abstract is a page listing a tongi and some pepeha. The tongi is a prophetic saying spoken by Tawhiao about the area where my marae (tribal meeting area) is located. The pepeha are tribal sayings that identifies where I am from. These serve not only to introduce myself as the author but also to accept responsibility and accountability for the research that I have undertaken in this thesis. No translation is given as the intended audience will understand what has been written.

The acknowledgements section, entitled He Tuku Aroha, has messages of gratitude written in many languages because I feel it is more appropriate that people are acknowledged in their own language, whether that is their first language, their indigenous language or the language they have the most affinity with. Again, no translations will be given as the intended recipients of this gratitude will understand what has been written.

This thesis does not have a section detailing my original contribution. As a Māori academic I am uncomfortable extolling my research. This thesis, and previous publications where I am listed as the primary author, are original contributions.

Although this document was written to conform to the requirements of a PhD thesis submitted at the University of Waikato, it was also written for a Māori audience. Because of this, certain writing conventions were used that allowed the writing style to be more comfortable to the Māori ear. These conventions included: not italicising reo Māori words (they are not foreign), the use of the phrase ‘te reo Māori’ or ‘reo Māori’ for Māori language and the use of the name Aotearoa for New Zealand.

He Tuku Aroha

Auē te mamae! Auē taukiri e!

Kei te ngau kino tonu te tarauma.

Kei taku Arikini. Takoto mai i te okiokinga o ngō mātua tūpuna,
te huihuinga o te kahurangi, te kāpunipuni o te hunga wairua.

Moe mai rā. Moe mai rā.

Kei ngā kaituhi o ngā niupepa tawhito, koutou kua tatū i te pūehū i ō koutou wā,
kua heke te werawera, kua rere te wai i ā koutou pene, kua tukuna ngā kupu ki ngā
perēhi. Nā reira kua whai oranga tonu ngō koutou whakaaro, kua whai mana tonu
ngā āhuatanga, ngā tikanga, ngā reo, a kui mā, a koro mā. E rau rangatira mā, moe
mai.

Kei aku nei tūpuna, tēnei te tuku aroha ki runga ki a koutou katoa. Moe mai rā.
Koutou ko Nan, ko Grandad, ko Uncle Jimmy mā, e noho tonu ana ā wairua, e
ārahina tonutia ana, e tiakina tonutia ana ahau. E tangi ana te ngākau mō koutou.
Koutou hoki tēnā, ko Dad, ko Ihaka mā, kei te noho mokemoke tonu te whenua
nei. Ko te tōtara nui kātahi anō ka hinga ki Taranaki, e Koro Ted, moe mai. Ko
koe hoki tēnā Nena Benton, moe mai.

Otirā, koutou katoa kua riro ki te kāpunipunitanga o ngā wairua, ki tua o te pae o
maumahara, takoto mai. Moe okioki atu rā. Kāti.

Me mihi ka tika ki tōku nei whānau, Mum, Dad, my mind still readily recalls how
you sold a cow so that I could first attend tertiary education. Your support and
aroha has always been unconditional. I will never be able to repay you and will
strive to follow your example in my paternal care of your mokopuna.

E taku whaea, Haupai, nāu ahau i ārahi, nāu ahau i poipoi, nāu ahau i whakatupu ki ngā āhuatanga me ngā taonga ngō tātou ake nei. Ko koe te whakatinanatanga o te kupu tautoko. Koutou hoki ko ngōku nei whaea ko Auntie Colleen, ko whaea Te Rita, ko whaea Terehia. Nā tō koutou tautoko kua whai kaha tēnei ki te takatū ake i āku nei mahi.

Kei ngā tuākana, tēina, tuāhine o te whānau Keegan, o te whānau Papesch, o te whānau Katipa hoki, otirā o ngā whānau whānui, kei te tuku aroha nui ki a koutou katoa. Tēnā rā koutou.

Professor Mark Apperley, thank you. It seems too much of a coincidence that one of your ancestors was an editor of one of the early Niupepa. You were one of those instrumental in starting the Niupepa online work and have been the most influential person on my career. Such was your influence on this research that when you took sabbatical leave I had to suspend this thesis. I will always be indebted and forever grateful.

I would also like to thank my other two supervisors Dr Richard Benton and Associate Professor Sally Jo Cunningham. Tēnā hoki kōrua. You have both been asked to undertake supervision work in very trying times. This thesis would not have been possible without your input and would have been immeasurably better if I had been able to undertake all that you two have suggested.

A special thanks to my proof readers. Rob Akscyn, Megan Symes and Haupai Puke. You all undertook separate but invaluable roles, with short notice and requests for quick turn-rounds. You all completed the proofing to a high degree of precision and accuracy. I am indebted. I regret if your advice has not been followed correctly.

I am very grateful to the Department of Computer Science at the University of Waikato for giving me the time, support and encouragement to undertake this work. In particular, Professor Mark Apperley, Associate Professor Steve Reeves, Associate Professor Geoff Holmes, and Associate Professor Tony McGregor, the various chairpersons of the Computer Science Department while I have been engaged on this research.

A special thanks to Stefan Boddie and the team at DL Consulting, and to Katherine Don and Michael Dewsnip of the NZDL, for making your precious time available to assist me when I needed you. Thanks also to Gareth Judsen for your excellent and ongoing work with the Perl scripts.

Makemake wau e mahalo aku i ka ‘ohana a me na hoa ma Hawai‘i, ia ‘oe Dr Robert Stauffer. E mahalo aku nei wau ia ‘oukou no kou kako‘o i ka mana‘o Niupepa me ka Nūpepa. A makemake wau e mahalo ia ‘oukou a pau no ka malama ana ia‘u i ka‘u mau hele ‘ana i Hawai‘i no ka ‘imi na‘auao.

E ha‘awi ana wau i ko‘u aloha i ka Hale Kuamo‘o a me keia mau po‘e, ‘o Keiki Kawai‘ae‘a lakou ‘o Dr Kalena Silva, ‘o Pila Wilson, ‘o Larry Kimura, ‘o Keola Donaghy, a me ka pua nani ma waena ‘o ‘oukou, ‘o Kiele Akana-Gooch. Mahalo a nui loa.

I would also like to thank all those people who agreed to be interviewed. Kei te mihi atu, mahalo nui, danke schön, 感谢你, muhuway su, dankie, thank you.

A special thanks also to Dr Gordon Paynter, Dr Maaka Laws, Dr Lachy Paterson and Dr James McLellan. Your theses provided ideas, encouragement and examples of which I have strived to follow.

Kei te tuku mihi atu hoki ki ngā kaiwhakawā. E te rangatira, e te whetū Māori o te kaupapa rorohiko, Maaka, tēnā koe. Ko tēnei te mihi atu ki a koe mō ngā mahi rorohiko katoa kua tutuki pai i a koe. Hei aha? Hei painga mō Ngāi Tātou. Tēnā koe kua whakawātea mai koe hei āta whakapae i te mahi rangahau nei.

Diolch i'r holwyr. Yr arweinydd parchus, yr goleuad yn yr iaith Gymraeg yn y maes cyfrifiadurol, Dr Daniel Cunliffe, rwy'n eich salwtio chi. Yr ydwyf yn eich diolch am bopeth y gallasoch ei wneud dros yr iaith Gymraeg yn y maes cyfrifiadurol. Bydd y gwaith hwn o fudd i genedlaethau i ddod. Yr ydwyf yn eich diolch am roi eich amser i'r archwilia hon gweithia. Diolch yn fawr.

Kei a koutou aku tamariki tokowhā, tōku ngākau. Ko koe tēnā Tamihana, koutou ko Tyler, ko Kaurukiruki, ko Ngāpera. E tangi ana te ngākau mō ngā haora i roto i ngā rā i roto i ngā pō i huri kē atu te aro i a koutou i huri kē mai ki te patopato i ngā kōrero o te tuhinga roa nei. E pupū ake ana he wai kei aku whatu i a au e tuhi ana i tēnei he nui nō tōku aroha mō koutou.

E tōku hoa rangatira, e Ā, kei a koe āku kupu whakamutunga. He waimarie nōku i te wā i tūtaki ngātahi ai tāua i Hiona. He waimarie nōku i te wā i mārena ngātahi ai tāua i Mangakoea. He waimārie nōku ia rā ka noho ngātahi ai tāua i a tāua. He nunui rawa te aroha, he aroha mutunga-kore. Nāu ahau i whakawātea kia rangahau ai kia tuhituhi ai i te tuhinga roa nei. Nāu te whare i whakarite, nāu ngā amuamu me ngā raruraru o te kāinga i whakapai, nāu ngā tamariki i whakatipu, nāu ngā kaputī me ngā pihikete i whōmai i waenganui pō, nāu ngā kupu akiaki i whakakaha mai i ngā uauatanga me ngā pōuritanga o tēnei mahi. Nā tāu mahi i whakaoti ai ahau te tuhinga roa nei. Me kī, nāhau te tuhinga nei, ā, kei te whōatu tēnei tuhinga roa māhau, me tōku arohanui hoki.

Table of Contents

Title	i
Tongi and Pepeha	iii
Abstract	v
Preface	vii
He Tuku Aroha	ix
Table of Contents	xiii
List of Figures	xxiii
List of Tables	xxvii
Abbreviations	xxxii
Chapter 1 Introduction	1
Why Investigate Digital Libraries?.....	2
Why Use the Term Indigenous language?.....	2
1.1 Thesis Statement.....	4
1.2 Background.....	4
1.2.1 The Decline of Te Reo Māori.....	5
1.2.2 Information Technology and Indigenous Languages ...	7
1.2.3 The Niupepa Digital Library	10

1.3	Thesis Approach	12
1.4	Thesis Outline	13
 Chapter 2 Transaction Log Analysis		 15
2.1	Introduction.....	15
2.2	Collecting Datasets	18
2.2.1	Dataset 1 NZDL 2005	19
2.2.2	Dataset 2 National and Dataset 3 International.....	19
2.2.3	Dataset 4 DefMi and Dataset 5 DefEn.....	20
2.2.4	Dataset 6 Hawaiian Nūpepa 2005.....	23
2.2.5	Summary of the Six Datasets of Web Logs	24
2.3	Preparing Datasets	25
2.3.1	Removing Redundant CGI Arguments.....	26
2.3.2	Removing Extraneous Requests.....	29
	Removing Non-Niuepa requests	29
	Removing Requests From Web Robots	30
	Removing Requests With Inappropriate Lang. Sett....	32
	Removing Requests With Other Incorrect Settings	32
2.3.3	Extraneous Requests Summary.....	33
2.4	Defining Sessions.....	33
2.4.1	Defining Unique Identifiers	34
	IP Address or Cookies/Z Argument?	35
2.4.2	Defining Session Types	36
2.4.3	Defining Session Language Types.....	37
2.4.4	Collating Sessions	40
2.5	Perl Scripts and Statistical Definitions	40
2.5.1	Apparent Reading	41
2.5.2	Language Switching Data	42
2.5.3	Determining Statistical Significance.....	42
2.6	Summary	43

Chapter 3	Reo Māori Usage of the Niupepa Digital Library	45
3.1	Niupepa Interface Description.....	46
	Accessing Niupepa Documents.....	48
	Niupepa Document Formats.....	51
3.2	Was Te Reo Māori Used by Niupepa Clients?.....	55
3.2.1	Was Te Reo Māori Used in all Session Types?.....	56
3.2.2	When Was Te Reo Māori Used?.....	58
3.3	Exploratory Behaviours.....	61
3.3.1	How Long Were Exploratory Sessions?.....	62
3.3.2	How Long to Read Interface Pages?.....	63
3.3.3	Which Interface Pages Were Requested?.....	65
3.3.4	Summary—Exploratory Sessions.....	66
3.4	Niupepa Interface Behaviours.....	67
3.4.1	How Long Were Usage Sessions?.....	68
	How Long in Time Were Usage Sessions?.....	69
	How Long in Number of Requests?.....	70
3.4.2	How Long to Read Interface Pages?.....	72
3.4.3	How Many Clients Returned?.....	73
3.4.4	Where Did Usage Sessions Originate?.....	74
	Originating Academic Institutions.....	76
	Originating Countries.....	78
3.4.5	Which Interface Pages Were Requested?.....	80
3.4.6	How Often Was the Interface Language Switched?...	81
3.4.7	When Was the Interface Language Switched?.....	82
3.4.8	Summary—Interface Behaviours.....	84
3.5	Document Access Behaviours.....	85
3.5.1	How Were Documents Accessed?.....	85
3.5.2	Which Document Formats Were Favoured?.....	87
3.5.3	Which Document Types Were Favoured?.....	88
3.5.4	Which Publication Years Were Favoured?.....	89
3.5.6	Documents With Associated Extracted Text Files.....	92
3.5.7	Documents With Associated English Abstracts.....	94
3.5.8	How Long to Read Documents?.....	95

3.5.9	Summary—Document Access Behaviours.....	96
3.6	Query Behaviours	97
3.6.1	How Many Queries Were Submitted?	97
3.6.2	How Many Query Terms Were Submitted?	98
3.6.3	What Topics Were Queried?.....	99
3.6.4	What Was the Language Of the Query Terms?	104
3.6.5	Were Query Parameters Altered?	105
3.6.6	Were Query Operators Used?	106
3.6.7	Documents Requested Following a Query.....	106
3.6.8	Formats Favoured Following a Query	107
3.6.9	Document Read Times Following Query	108
3.6.10	Summary—Query Behaviours.....	109
3.7	Te Reo Māori Usage/English Language Usage.....	110
	Behaviour differences in exploratory sessions.....	110
	Behaviour differences in usage sessions.....	111
3.8	Bilingual Usage.....	113
3.9	Summary.....	115

Chapter 4 Comparing International and National Reo Māori Usage 119

4.1	Was Te Reo Māori Used by International Clients?	120
4.1.1	Did International Clients Use the Niupepa?	120
4.1.2	Did International Clients Show Interest?	122
4.1.3	Did International Clients Use Te Reo Māori?	123
4.1.4	When Was Te Reo Māori Used?.....	124
4.1.5	Summary—Was Te Reo Māori Used?.....	126
4.2	International and National Exploratory Behaviours	127
4.2.1	How Long Were Exploratory Sessions?	128
4.2.2	How Long to Read Interface Pages?.....	129
4.2.3	Which Interface Pages Were Requested?	129
4.2.4	Summary—Exploratory Sessions	130
4.3	International and National Interface Behaviours.....	131
4.3.1	How Long Were Usage Sessions?	132

	How Long in Time Were Usage Sessions?	132
	How Long in Number of Requests?	134
4.3.2	How Long to Read Interface Pages?	136
4.3.3	How Many Clients Returned?	137
4.3.4	Which Interface Pages Were Requested?	138
4.3.5	How Often Was the Language Switched?	139
4.3.6	Summary—Interface Behaviours	140
4.4	International and National Document Access Behaviours	141
4.4.1	How Were Documents Accessed?	141
4.4.2	Which Document Formats Were Favoured?	143
4.4.3	Which Document Types Were Favoured?	144
4.4.4	How Long to Read Documents?	146
4.4.5	Summary—Document Access Behaviours	147
4.5	International and National Query Behaviours	148
4.5.1	How Many Queries Were Submitted?	148
4.5.2	How Many Query Terms Were Submitted?	149
4.5.3	What Topics Were Queried?	150
4.5.4	What Was the Language Of the Query Terms?	154
4.5.5	Were Query Parameters Altered?	154
4.5.6	Were Query Operators Used?	155
4.5.7	Documents Requested Following a Query	156
4.5.8	Formats Favoured Following a Query	156
4.5.9	Document Read Times Following a Query	157
4.5.10	Summary—Query Behaviours	158
4.6	Te Reo Māori Usage/English Language Usage	159
4.6.1	Interface Behaviour Differences	160
4.6.2	Document Access Behaviour Differences	160
4.6.3	Query Behaviour Differences	161
4.6.4	Summary—Behaviour Differences	162
4.7	Summary	163
4.7.1	Discussion	164

Chapter 5	The Effect of the Default Language Setting On Te Reo Māori Usage	167
5.1	Did Te Reo Māori Usage Increase?	168
5.1.1	Effect on the Number of Niupepa Requests.....	168
5.1.2	Were Potential Niupepa Clients Discouraged?.....	169
5.1.3	Effect on Number of Te Reo Māori Requests.....	171
5.1.4	Effect on Number of Te Reo Māori Clients.....	172
5.1.5	Effect on When Usage Occurred?.....	173
5.1.6	Summary—Effect on Te Reo Māori Usage	175
5.2	Effect on Exploratory Behaviours	176
5.2.1	Effect on Interface Page Read Times.....	177
5.2.2	Effect on Which Interface Pages Were Requested ...	180
5.2.3	Summary—Effect on Exploratory Behaviours.....	181
5.3	Effect on Interface Behaviours	182
5.3.1	Effect on Usage Session Length	183
	Usage Session Duration in Time.....	184
	Usage Sessions Length in Number of Requests.....	185
5.3.2	Effect on Time to Read Interface Pages.....	187
5.3.3	Effect on Return Requests.....	189
5.3.4	Effect on Which Interface Pages Were Requested ...	189
5.3.5	Effect on Interface Language Switching.....	190
5.3.6	Summary—Effect on Interface Behaviours.....	192
5.4	Effect on Document Access Behaviours.....	193
5.4.1	Effect on How Documents Were Accessed	194
5.4.2	Effect on Document Formats Requested	195
5.4.3	Effect on Which Document Types Were Requested	197
5.4.4	Effect on Reading Time Of Documents.....	199
5.4.5	Summary—Effect on Document Access	199
5.5	Effect on Query Behaviours.....	200
5.5.1	Effect on Number of Queries	201
5.5.2	Effect on Number of Query Terms	202
5.5.3	Effect on Language of Query Terms.....	203
5.5.4	Effect on Altering of Query Parameters	204
5.5.5	Effect on Use of Query Operators.....	205

5.5.6	Summary—Effect on Query Behaviours.....	205
5.6	Effect on Use in English.....	206
5.7	Summary.....	208
5.7.1	Discussion.....	210

Chapter 6 Hawaiian Language Usage of the Nūpepa Digital Library 211

6.1	The Hawaiian Nūpepa Digital Library.....	213
6.1.1	Capturing the Hawaiian Nūpepa.....	213
6.1.2	Hawaiian Nūpepa Online.....	214
6.1.3	Hawaiian Nūpepa Interface Description.....	217
6.1.4	Nūpepa/Niupepa Interface Differences.....	221
6.2	Was the Hawaiian Language Used?.....	224
6.2.1	Was Hawaiian Used in All Three Session Types? ...	225
6.2.2	Was the Hawaiian DL Used Bilingually?.....	227
6.2.3	When Was the Hawaiian Language Used?.....	227
6.2.4	Summary—Was the Hawaiian Language Used?.....	229
6.3	Exploratory Behaviours.....	230
6.3.1	How Long Were Exploratory Sessions?.....	230
6.3.2	How Long to Read Interface Pages?.....	232
6.3.3	Summary—Exploratory Behaviours.....	233
6.4	Nūpepa Interface Behaviours.....	234
6.4.1	How Long Were Usage Sessions?.....	235
	How Long in Time?.....	236
	How Long in Number of Requests?.....	237
6.4.2	How Long to Read Interface Pages?.....	238
6.4.3	How Many Clients Returned?.....	239
6.4.4	Where Did Usage Sessions Originate?.....	240
6.4.5	How Often Was the Interface Language Switched?.	242
6.4.6	When Was the Interface Language Switched?.....	243
6.4.7	Summary—Interface Behaviours.....	244
6.5	Document Access Behaviours.....	245
6.5.1	How Were Documents Accessed?.....	246

6.5.2	Which Document Formats Were Favoured?.....	247
6.5.3	Which Nūpepa Periodicals Were Favoured?	248
6.5.4	Which Publication Years Were Favoured?.....	251
6.5.5	Documents With Associated Extracted Text Files ...	254
6.5.6	Was the Detach Option Used?	255
6.5.7	How Long to Read Documents?	255
6.5.8	Summary—Document Access Behaviours.....	256
6.6	Query Behaviours	258
6.6.1	How Many Queries Were Submitted?	258
6.6.2	How Many Query Terms Were Submitted?	260
6.6.3	What Was The Language Of The Query Terms?	261
6.6.4	Were Query Parameters Altered?	262
6.6.5	Were Query Operators Used?	262
6.6.6	Documents Requested Following a Query.....	263
6.6.7	Formats Favoured Following a Query	264
6.6.8	Document Read Times Following a Query.....	265
6.6.9	Summary—Query Behaviours.....	266
6.7	Hawaiian Language Usage/English Language Usage	267
	Behaviour Differences in Exploratory Sessions	267
	Behaviour Differences in Usage Sessions	268
6.8	Bilingual Usage.....	271
6.9	Summary	273

Chapter 7 Why was the Indigenous Language Used by Digital Library Clients? 277

7.1	Web Questionnaire Methodology	278
7.1.1	Web Questionnaire Description	278
7.1.2	Web Questionnaire Presentation.....	280
7.2	Web Questionnaire Responses.....	281
7.2.1	What Did Clients Use the Digital Library For?	282
7.2.2	Which Research Sectors Used the Digital Library?..	284
7.2.3	What Did Clients Find Most Rewarding?.....	285
7.2.4	What Did Clients Find Most Frustrating?.....	287

7.2.5	What Language Was Used by Clients?	288
7.2.6	Further Comments From the Web Questionnaire.....	289
7.2.7	Summary—Web Questionnaire Responses.....	290
7.3	Client Interview Methodology	291
7.3.1	Client Interview Procedure.....	291
7.3.2	Finding Clients To Be Interviewed	293
7.4	Client Interview Responses	296
7.4.1	Interviews With UsgMi Clients.....	296
	UsgMi Interview #1.....	297
	UsgMi Interview #2.....	298
	UsgMi Interview #3.....	299
7.4.2	Interviews With UsgHaw Clients	300
	UsgHaw Interview #1	300
	UsgHaw Interview #2.....	302
	UsgHaw Interview #3	303
	UsgHaw Interview #4.....	304
	UsgHaw Interview #5.....	305
7.4.3	Interviews With UsgEn Clients	306
	UsgEn Interview #1	307
	UsgEn Interview #2.....	307
	UsgEn Interview #3	308
	UsgEn Interview #4	309
	UsgEn Interview #5	310
	UsgEn Interview #6	311
	UsgEn Interview #7	312
	UsgEn Interview #8.....	313
7.4.4	Client Interviews Summary	314
	UsgMi Client Interviews	315
	UsgHaw Client Interviews.....	316
	UsgEn Client Interviews.....	317
7.5	Summary.....	319

Chapter 8 Research Findings	323
8.1 Reo Māori Usage	324
8.2 International Reo Māori Usage	325
8.3 Default Language Effect	326
8.4 Hawaiian Language Usage	327
8.5 Indigenous and non-Indigenous Usage	328
8.6 Bilingual Users	330
8.7 Interface Usability and Design	331
Translations	331
Content Description and Format	332
Query Activity	333
Navigation and Layout	334
8.8 Results Requiring Further Investigation	334
8.9 Claims Revisited	336
8.10 Final Comments	337
References	339
Appendix A Perl Script Summary	345
Appendix B Hawaiian Nūpepa Query Terms	355
Appendix C Interview Forms	357

List of Figures

Figure 3-1: The home page of the Niupepa web site (reo Māori interface).....	47
Figure 3-2: The Preferences page of the Niupepa web site (English interface). .	48
Figure 3-3: Query results page of the Niupepa web site (English interface).....	49
Figure 3-4: Series list of the Niupepa web site (English interface).	49
Figure 3-5: The date list of the Niupepa web site (English interface).	50
Figure 3-6: The Extracted Text document of a newspaper page.	51
Figure 3-7: The Preview image of a newspaper page.....	52
Figure 3-8: The Full Size image of a newspaper page.....	52
Figure 3-9: An example of an English Abstract on the Niupepa web site.....	53
Figure 3-10: A Bibliographic Commentary on the Niupepa web site.	54
Figure 3-11: Language of Requests to the Niupepa Digital Library.....	55
Figure 3-12: Requests to the Niupepa web site in 2005 divided into session types: single requests, exploratory sessions and usage sessions. ...	56
Figure 3-13: Comparing the interface language of the Niupepa web site in 2005 for single request, exploratory and usage sessions.	57
Figure 3-14: Comparing requests recorded per hour of the day in reo Māori sessions and English language sessions.	59
Figure 3-15: Comparing requests recorded per day of the week in reo Māori sessions and English language sessions.	59
Figure 3-16: Comparing requests recorded per month in reo Māori sessions and English language sessions.....	60
Figure 3-17: Usage Session requests by language type.	68
Figure 3-18: Comparing the durations of reo Māori usage sessions with English language and bilingual usage sessions.	69
Figure 3-19: Comparing the number of requests in reo Māori usage sessions with English language and bilingual usage sessions.	71
Figure 3-20: Comparing the origin of the 100 most active IP Addresses in usgMi, usgEn and usgBi sessions as indicated by domain name. .	75
Figure 3-21: Comparing request percentages from academic institutes for top 100 IP Addresses, grouped by usgMi, usgEn and usgBi.....	77

Figure 3-22: Number of unique IP Addresses submitting 20 or more requests from overseas countries.	78
Figure 3-23: Comparing where in the session the language was switched from English to reo Māori for usgMi, usgEn and usgBi sessions.	82
Figure 3-24: Comparing where in the session the language is switched from reo Māori to English for usgMi, usgEn and usgBi sessions.	83
Figure 3-25: Comparison of document types requested in top 100 documents by unique user for usgMi, usgEn and usgBi sessions.	89
Figure 3-26: Requests for documents by all clients per publication year.	91
Figure 3-27: Percentage of requested documents in selected years by usgMi clients and usgEn clients.	92
Figure 3-28: The number of queries submitted per usage session for usgMi, usgEn and usgBi sessions.	98
Figure 3-29: Topics of 100 most requested query terms entered by usgMi clients, usgEn clients and usgBi clients.	103
Figure 3-30: Language of query terms entered by usgMi, usgEn and usgBi clients.	104
Figure 4-1: National and International Niupepa requests.	121
Figure 4-2: Comparison of national and international requests in session types; single requests, exploratory sessions & usage sessions. ...	123
Figure 4-3: Comparing the national and international interface languages settings for single request, exploratory and usage sessions.	124
Figure 4-4: Comparing requests recorded per month in national and international usgMi sessions.	125
Figure 4-5: International and national comparisons of top 100 document types requested by unique users in usgMi and usgEn sessions.	145
Figure 4-6: International and national comparison of query term topics in top 100 query terms submitted.	153
Figure 5-1: Default Language Setting for requests in 2005.	169
Figure 5-2: Comparison of defMi and defEn requests in single request, exploratory and usage sessions.	170
Figure 5-3: Comparing language use in sessions when default language was set to reo Māori (defMi) and English (defEn).	172
Figure 5-4: Comparing usage requests per hour with the default web site language set to te reo Māori (defMi) and English (defEn).	174
Figure 5-5: Comparing usage requests per day with the default web site language set to reo Māori (defMi) and English (defEn).	174
Figure 5-6: Comparing usage requests per month with the default web site language set to reo Māori (defMi) and English (defEn).	175

Figure 5-7: defMi and defEn comparisons of top 100 page types requested by unique users in usgMi and usgEn sessions.....	198
Figure 6-1: The home page of the Nūpepa web site (Hawaiian interface).	218
Figure 6-2: The results page (English interface) after a query for ‘nupepa’	219
Figure 6-3: A Nūpepa extracted text page (English interface).	219
Figure 6-4: Series list of the Nūpepa web site (English interface).	220
Figure 6-5: Date list of the Nūpepa web site (English interface).....	221
Figure 6-6: Language of Requests to the Nūpepa Digital Library.....	224
Figure 6-7: Requests to the Nūpepa web site in 2005 divided into session types: single requests, exploratory sessions and usage sessions.	225
Figure 6-8: Comparing the interface language of the Nūpepa web site for single request sessions, exploratory sessions and usage session.	226
Figure 6-9: Comparing requests recorded per hour of the day by Nūpepa clients in Hawaiian language, English and bilingual sessions. ...	228
Figure 6-10: Comparing requests recorded per day of the week by Nūpepa clients in Hawaiian language, English and bilingual sessions. ...	228
Figure 6-11: Comparing requests recorded per month by Nūpepa clients in Hawaiian language, English and bilingual sessions.	229
Figure 6-12: Usage Session requests by language type.	235
Figure 6-13: Comparing the durations of Hawaiian language usage sessions with English language and bilingual usage sessions.	236
Figure 6-14: Comparing the number of requests in Hawaiian language usage sessions with English and bilingual usage sessions.	237
Figure 6-15: Comparing the origin of the 100 most active IP addresses in usgHaw, usgEn and usgBi	241
Figure 6-16: Comparing where the interface language was switched from Hawaiian to English for usgHaw, usgEn and usgBi sessions.	243
Figure 6-17: Comparing where the interface language was switched from English to Hawaiian for usgHaw, usgEn and usgBi sessions.	244
Figure 6-18: Periodicals requested by usgHaw, usgEn and usgBi clients.....	250
Figure 6-19: Number of requests for identified documents per publication year.	252
Figure 6-20: Percentage of requested documents in selected publication years by usgHaw, usgEn and usgBi clients.	253
Figure 6-21: Comparisons of number of queries per session in usage sessions.	259
Figure 6-22: Language of query terms entered by usgHaw, usgEn and usgBi clients.....	261
Figure 7-1: A screen shot of the web questionnaire form.....	279

Figure 7-2: Comparison of usgMi and usgEn responses to the question on why clients use the Niupepa collection.....	282
Figure 7-3: Responses by usgMi and usgEn clients to the question on which research sectors were using the Niupepa collection	284
Figure C-1: Bill of Rights for Research Participants	358
Figure C-2: Research Consent Form.....	359

List of Tables

Table 2-1: The dates, times and (new) default languages that were set on the Niupepa web site in 2005	22
Table 2-2: Information about web log datasets listing descriptions, web server, when recorded, requests and dataset file size	25
Table 2-3: Number of requests removed from initial raw log file Dataset 1 to create the prepared log file.	33
Table 3-1: Comparisons of length (in seconds) of expMi and expEn sessions. ..	62
Table 3-2: Comparisons of number of requests in expMi and expEn sessions. ..	62
Table 3-3: expMi and expEn apparent read times of Niupepa interface pages in exploratory sessions.	64
Table 3-4: Statistical summaries comparing apparent read times in expMi and expEn exploratory sessions.	64
Table 3-5: The numbers and percentages of web site interface pages requested in exploratory sessions by expMi and expEn.	65
Table 3-6: Summary statistics comparison of usgMi, usgEn, usgBi and total apparent page read times of interface pages in usage sessions.	72
Table 3-7: The numbers and percentages of interface pages requested by usgMi, usgEn and usgBi clients.	80
Table 3-8: Counts and percentages comparing how documents were accessed by usgMi, usgEn, usgBi clients and total usage session clients.	86
Table 3-9: Document formats requested by usgMi, usgEn and usgBi clients.	87
Table 3-10: Comparing the number of requests for documents with and without extracted text files in usgMi and usgEn usage sessions. ...	93
Table 3-11: Comparing average apparent read times of document types in usgMi, usgEn and usgBi usage sessions.....	95
Table 3-12: The number of query terms submitted per query in usgMi, usgEn, usgBi and the total usage sessions.....	99
Table 3-13: The query terms (with counts) most commonly submitted by usgMi, usgEn and usgBi clients.	101
Table 3-14: Comparing formats requested not involving a query with those following a query from usgMi, usgEn and usgBi clients.	107
Table 3-15: Comparing apparent read times of non-query documents with query documents by usgMi, usgEn and usgBi clients.....	108

Table 4-1: Number and percentages of web site interface pages requested in international and national exploratory sessions.	130
Table 4-2: International and national comparisons for duration groupings of usgMi, usgEn and usgBi sessions showing chi square p values. .	133
Table 4-3: International and national comparisons of statistical summaries for durations (in minutes) of usgMi, usgEn and usgBi sessions.	134
Table 4-4: International and national comparisons of request groupings in usgMi, usgEn and usgBi sessions with chi square p values.	135
Table 4-5: International and national comparisons of statistical summaries for length (in requests) of usgMi, usgEn and usgBi usage sessions..	135
Table 4-6: International and national comparisons of apparent read times of Niupepa interface pages in usgMi, usgEn and usgBi sessions. ...	137
Table 4-7: International and national comparison of percentages of Niupepa web site interface pages requested in usage sessions.....	139
Table 4-8: Comparing national and international switching counts and percentages for usgMi, usgEn and usgBi sessions.....	139
Table 4-9: International and national comparison by percentage of how documents were requested by usgMi, usgEn and usgBi clients. .	142
Table 4-10: International and national comparisons of document types requested by usgMi, usgEn and usgBi clients.	143
Table 4-11: International and national average apparent read times (in seconds) of documents in usgMi, usgEn and usgBi sessions.	146
Table 4-12: International and national comparisons of number of sessions that contain queries in usgMi, usgEn and usgBi sessions.....	149
Table 4-13: International and national percentage comparisons for the number of terms in a query for usgMi, usgEn and usgBi sessions.	150
Table 4-14: The query terms (with counts) most commonly submitted by international and national clients.	151
Table 4-15: International and national percentages of document formats requested following a query.....	157
Table 4-16: Apparent read times (seconds) by international and national usgMi clients following the submission of a query.	158
Table 5-1: defMi and defEn apparent read times of interface pages in expMi and expEn sessions.	178
Table 5-2: Statistical summaries comparing apparent read times in expMi and expEn sessions with defMi and defEn settings.....	179
Table 5-3: Numbers and percentages of interface pages requested by expMi and expEn clients with defMi and defEn language settings.	180
Table 5-4: defMi and defEn comparisons for length of session groupings (in minutes) of usage sessions.	184

Table 5-5: defMi and defEn statistical summaries for duration (in minutes) of usgMi, usgEn and usgBi usage sessions.....	185
Table 5-6: defMi and defEn request groupings with p values.	186
Table 5-7: defMi and defEn comparisons of statistical summaries for length (in requests) of usgMi, usgEn and usgBi usage sessions.	187
Table 5-8: defMi and defEn statistical comparisons of apparent read times of Niupepa interface pages in usgMi, usgEn and usgBi sessions....	188
Table 5-9: defMi and defEn comparison of percentages of Niupepa web site interface pages requested in usgMi, usgEn and usgBi sessions. .	190
Table 5-10: defMi and defEn total interface switches, switches per week and total switches/total requests for all usage sessions types.....	191
Table 5-11: defMi and defEn switches per session for all usage types.	192
Table 5-12: defMi and defEn comparison by percentage of how documents were requested by usgMi, usgEn and usgBi clients.	194
Table 5-13: defMi and defEn comparisons of document types requested by usgMi, usgEn and usgBi clients.	196
Table 5-14: defMi and defEn comparisons of average apparent read times (in seconds) of documents by usgMi, usgEn and usgBi clients.....	199
Table 5-15: defMi and defEn comparisons of number of sessions that contained queries by usgMi, usgEn and usgBi clients.	201
Table 5-16: defMi and defEn percentage comparisons for the number of terms in a query in usgMi, usgEn and usgBi sessions.....	202
Table 5-17: defMi and defEn percentage comparisons for language of query terms submitted by usgMi, usgEn and usgBi clients.....	204
Table 6-1: Comparison of Nūpepa and Niupepa interface features as available to clients in 2005.	222
Table 6-2: Comparisons of lengths (in seconds) of expHaw and expEn sessions.	231
Table 6-3: Comparisons of number of requests in expHaw and expEn sessions.....	231
Table 6-4: expHaw and expEn apparent read times of Nūpepa interface pages in exploratory sessions.	232
Table 6-5: Summary statistics of apparent read times of interface pages in usgHaw, usgEn and usgBi sessions.....	238
Table 6-6: Counts and percentages comparing how documents were accessed by usgHaw, usgEn and usgBi clients.	246
Table 6-7: Document formats requested by usgHaw, usgEn and usgBi clients.	247
Table 6-8: Number of requests per Nūpepa periodical.	249
Table 6-9: The number of requests for documents with and without associated extracted text files by clients, usgEn and usgBi clients.....	254

Table 6-10: Average apparent read times of document types by usgHaw, usgEn and usgBi clients.....	256
Table 6-11: Queries per session in usgHaw, usgEn and usgBi sessions.....	259
Table 6-12: The number of query terms submitted per query by usgHaw, usgEn and usgBi clients.....	260
Table 6-13: Comparing documents requested not involving a query with those following a query from usgHaw, usgEn and usgBi clients.....	264
Table 6-14: Comparing apparent read times not involving a query with read times following a query for usgHaw, usgEn and usgBi clients. ..	265
Table 7-1: Summary information about the clients who were interviewed using the Niupepa or Nūpepa digital libraries	295
Table A-1: Summary of script process_raw_logs.pl	346
Table A-2: Summary of create-sessions.pl.....	346
Table A-3: Summary of split-sess-lang.pl	347
Table A-4: Summary of analyse-single-types.pl	347
Table A-5: Summary of session-stats.pl.....	348
Table A-6: Summary of logs.pl	349
Table A-7: Summary of analysis.pl.....	349
Table A-8: Summary of gui.pl.....	350
Table A-9: Summary of renameFiles.pl	351
Table A-10: Summary of renameLengths.pl	352
Table B-1: Query terms submitted by usgHaw, usgEn and usgBi clients to the Hawaiian Nūpepa digital library.....	356

Abbreviations

CGI	Common Gateway Interface
HTTP	Hypertext Transfer Protocol
IP Address	Internet Protocol Address
NZDL	New Zealand Digital Library
OCR	Optical Character Recognition
TLA	Transaction Log Analysis
Web	World Wide Web
defEn	default English; English was the default language of the web site interface when the client requested the home page.
defHaw	default Hawaiian; Hawaiian was the default language of the web site interface when the client requested the home page.
defMi	default reo Māori; te reo Māori was the default language of the web site interface when the client requested the home page.
expEn	exploratory English; an exploratory session where the language of the interface was mostly set to English (see §2.4.3 for a precise definition of exploratory sessions).
expHaw	exploratory Hawaiian; an exploratory session where the language of the interface was mostly set to Hawaiian
expMi	exploratory reo Māori; an exploratory session where the language of the interface is mostly set to te reo Māori

- usgEn usage English; a usage session where the language of the interface was mostly set to English (see §2.4.3 for a precise definition of usage sessions).
- usgHaw usage Hawaiian; a usage session where the language of the interface was mostly set to Hawaiian.
- usgMi usage reo Māori; a usage session where the language of the interface was mostly set to te reo Māori.
- usgBi a usage session where the language of the interface was set to both English and the indigenous language for a significant number of the requests.

1 Introduction

‘Tōku reo, tōku ohooho
Tōku reo, tōku māpihi maurea
Tōku reo, tōku whakakai marihi
Aue! Aue! Aue! Te mamae ki taku taonga e!

My language, my cherished possession
My language, my object of affection
My language, my precious adornment
Aue! Aue! Aue! What pain I feel at the loss
of something so valuable. (Kāretu 1996)ⁱ

These words describe the importance of te reo Māoriⁱⁱ to the Māori people. They are words from a whakawātea,ⁱⁱⁱ titled, *He Aha Rā Kei Te Tau o Taku Ate?* (What Is This That is Tearing Me Apart?).^{iv} The whakawātea raises the serious question of how te reo Māori can survive, and challenges Māori to ‘get your act together’ so that the language can become revitalised again.

The research described in this thesis suggests that the revitalisation of te reo Māori and other indigenous languages can be assisted by their use in digital libraries.

ⁱ See Kāretu (1996, pp. 75–76).

ⁱⁱ The Māori language.

ⁱⁱⁱ An item performed in kapahaka (performing arts group) competitions where the performers exit the stage. Whakawātea literally means to leave clear.

^{iv} This whakawātea was composed by Tīmoti Kāretu and first performed in 1988 by the Te Whare Wānanga o Waikato kapahaka at national competitions in Whāngārei.

Why Investigate Digital Libraries?

Communication is increasingly occurring and information is increasingly being provided, electronically. From the advent of the radio and telephone, through the widespread use of television and video, to the propagation of computers and the Internet, the media by which societies communicate and transmit information have taken many electronic forms. Languages are likely to survive and prosper if they can be used in these new electronic environments.

One particular electronic environment with the potential to provide new situations where languages can be used, and thus assist in language revitalisation, is a digital library. Witten and Bainbridge (2002, pp xxvi) define a digital library as a focused collections of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organisation and maintenance. A digital library's potential to assist use of a language is twofold. First it has the potential to make available a large amount of information about a language and/or in a language readily available to speakers of that language. With suitable content in a particular language it has the potential to attract a large number of those language speakers. Second, it has the potential to offer an interface in any written language, thus offering speakers of that language an opportunity to use their language in a modern electronic environment.

If a digital library with an indigenous language interface makes available indigenous language material, and that material is highly sought after, then there is the potential for the digital library to be *used* in the indigenous language. Revitalisation of indigenous languages is likely to occur if indigenous language speakers choose to use their indigenous languages in new environments such as digital libraries.

Why Use the Term Indigenous Language?

Indigenous language is a phrase used in this thesis to signify the traditional language that is spoken by indigenous people, for example te reo Māori in

Aotearoa/New Zealand.^v Indigenous languages are languages that originate in a particular country or locale and are usually spoken by people who have an affinity to that locale. While many languages could be classified as indigenous languages, this research uses the term indigenous language to describe traditional languages that are not in a dominant position and are in threat of extinction. The precarious survival of these languages is the main reason that this research is so important.

The decision to use the term indigenous is not taken lightly as there are an abundance of terms that have been used to describe similar language research in the computing and internet environments. These terms include native, aboriginal, heritage, local, community, non-state, endangered, threatened, disadvantaged, lesser-used, less commonly taught and minority. While the language primarily studied in this research, te reo Māori, could be described using all of these adjectives, the term indigenous is preferred as it does not imply a language from a specific timeframe and it does not depict a language in a marginalised or negative manner. The term indigenous does portray a language with mana,^{vi} characteristically mana whenua^{vii} and is the term that is preferred in recent publications by Māori scholars (Durie, 2004; Smith, 1999; Walker, 2005).

Finally, the term indigenous language is used as a reminder to the reader that indigenous language users are, in the most part, indigenous users. Their purposes, their methods and their characteristics in using their indigenous language are indigenous purposes, indigenous methods and indigenous characteristics. This analysis of indigenous language usage by clients^{viii} of digital libraries will take into account indigenous perspectives and indigenous viewpoints.

^v Aotearoa/New Zealand is a modern phrase used to describe the country commonly known as New Zealand. For the sake of brevity Aotearoa/New Zealand will be reduced to Aotearoa.

^{vi} Mana: prestige, authority, held in great esteem.

^{vii} Mana Whenua: in this case mana that is derived from having a long existence in or long association with a certain locale.

^{viii} Client is the term used in this research to describe users who log on to a digital library and use it through the web interface.

1.1 Thesis Statement

This thesis charts indigenous language usage by clients of a digital library and demonstrates how digital libraries can provide an environment to support the use of indigenous languages.

This claim is that digital libraries provide a modern electronic medium where the use by choice, that is use without coercion or encouragement of an indigenous language can occur. The claim is substantiated in two parts. The first explains *how* indigenous languages are used by clients of digital libraries and what settings encourage indigenous language use. The second part suggests *why* indigenous languages are used by clients of digital libraries and gives characteristics of digital library environments that promote the usage of the indigenous language.

1.2 Background

The indigenous language that receives the most scrutiny in this research is te reo Māori of Aotearoa/New Zealand. A brief description is provided of the origin of the word Māori, the decline and rise of the Māori population and te reo Māori use is outlined. This is followed by a summary of why information technology, including digital libraries, is so important to indigenous languages. The digital library comprehensively analysed in this research is the Niupepa^{ix} digital library. This section concludes by introducing the Niupepa digital library and explaining why this particular digital library is appropriate for an investigation of indigenous language use.

^{ix} Literally ‘the Newspapers’, a collection of newspapers published in Māori or for a Māori audience between 1840 and 1933. For a full description see §1.2.3 and §3.1.1.

1.2.1 The Decline of Te Reo Māori

Te reo Māori is threatened with extinction (Benton, 1978; Te Puni Kōkiri, 1998). The indigenous language of Aotearoa is a minority language amongst its indigenous people. This sub-section summarises the decline of te reo Māori emphasising the importance of this research in seeking ways to assist in its revitalisation.

The word Māori is commonly used to describe the indigenous people of Aotearoa.^x Traditionally the term māori means ‘usual, ordinary or normal’ (Williams, 1985). The phrase tāngata māori^{xi} was initially used by Māori to distinguish themselves from the early European explorers to Aotearoa. Later, European settlers adopted the shorter version of the term, Māori, to describe the indigenous people of Aotearoa and over time this term became globally accepted (Metge, 1967).

Te reo Māori has been classified as being in the Eastern-Polynesian group of languages (Harlow, 1987). It resides in the Tahitic sub-class of Central Eastern Polynesian languages, along with Austral (French Polynesia), Penryn (Cook Islands), Rakahanga-Manihiki (Cook Islands), Rarotongan (Cook Islands), Tahitian (French Polynesia), Tuamotuan (French Polynesia) and Rapa (Easter Islands) (Gordon, 2005).

Estimates suggest the Māori population numbered over 100,000 prior to European contact and then declined to as low as 42,000 in the late 1800’s, primarily due to introduced diseases and land alienation (Pool, 1991). As the Māori population decreased the European population increased. In 1840 the Māori population outnumbered the European population by 40:1 but by the end of the 19th century the European population outnumbered the Māori population by 16:1. From the 1900s the Māori population decline ceased and from the 1920s a gradual regeneration began because of an improved immunity to diseases and an

^x The indigenous people of the Cook Islands are also referred to as Māori, though more commonly as Cook Island Māori in Aotearoa.

^{xi} Tāngata is people, māori is normal or common; hence tāngata māori means the normal people.

improvement in social economic status (Pool, 1991). Results from the 2006 census indicate that there were 565,329 Māori living in Aotearoa (Statistics New Zealand, 2006). The 2006 census also showed that 14.6% of Aotearoa's resident population were Māori, 67.6% were European, and the remaining 17.8% consisted of other ethnicities.

The use of te reo Māori in Aotearoa declined with the decline in Māori population in the 19th century. A lower proportion of Māori people meant there was less need for the European settlers to learn te reo Māori. The passing and rigorous enforcing of the Native Schools Act in 1867, which emphasised the use of English and excluded the use of te reo Māori in the classroom and school grounds, was a contributing factor for the language shift that was to occur in Māori communities. The ban on te reo Māori in schools, together with the rapid urbanisation of Māori in the 1940s and 1950s and the pepper potting policy (Benton, 1997), led to a massive language shift to English amongst Māori people (Te Puni Kōkiri, 2004). While pockets of rural communities still existed where te reo Māori was the prominent language, research undertaken in the late 1970s suggested that reo Māori was destined to vanish in the near future (Benton, 1978, 1981).

The imminent threat to te reo Māori gave rise to many initiatives seeking to redress its impending extinction. Such initiatives included the establishment of Te Kōhanga Reo (The Language Nest—te reo Māori pre-schools) in 1982; Kura Kaupapa Māori (Reo Māori primary schools based on Māori ideology) in 1985; Te Taura Whiri i te Reo Māori (Māori Language Commission) in 1987; Te Māngai Pāho (Māori Broadcasting Agency) in 1993, and a host of other initiatives^{xii} and efforts by both community and government agencies.

While the revitalisation efforts gave rise to a lot of optimism, a large national reo Māori survey undertaken in 1995 indicated that te reo Māori was still in a critical state. Only 21.7% of Māori had a fluency above a low level (Te Puni Kōkiri, 1998). Recent figures from the 2006 census list 131,613 Māori (24.4% of Māori) who indicated that they could speak te reo Māori (Statistics New Zealand, 2006).

^{xii} For a general description of Māori revitalisation efforts see Reedy (2000).

However, these figures do not indicate the complexity of reo Māori that is being spoken, the ability to read and write in Māori, or if intergenerational transmission is occurring. Intergenerational transmission is recognised as being crucial for language survival (Fishman, 1991). The outlook for the survival of te reo Māori as a language of everyday use is by no means assured.

To be revitalised as a language of everyday use, te reo Māori must be used by choice, without coercion or pressure, by its speakers in everyday situations. Opportunities in all endeavours of life must be created where the discretionary use of te reo Māori will occur. These include but are not limited to domestic settings, social events, sporting activities, education and business discourse, and all other opportunities both verbal and non-verbal when language is used for communication. One particular endeavour, a component of which is investigated in this thesis, is the use of language in information technology.

1.2.2 Information Technology & Indigenous Languages

The importance of information technology to indigenous languages has been described by Hale Kuamo‘o’s^{xiii} current director Keiki Kawai‘ae‘a, who states, “...In order for Hawaiian to feel like a real living language, like English, it needs to be seen, heard and utilized everywhere, and that includes the use of computers” (cited in Warschauer & Donaghy, 1997, pp. 352-353).

There are many ways that information technology, in the form of computing and networking devices, can be used to assist languages threatened with extinction. Information technology can be used to document and archive language information; create a wide range of multi-media resources for language learning (and preservation); connect communities of language speakers and thus create new environments where the threatened language can be used; as a source of new identity for speakers and to instil further pride in the language. David Crystal, in

^{xiii} A detailed description of Hale Kuamo‘o is given in §6.1.2.

his book *Language Death* discusses language revitalisation factors and postulates "An endangered language will progress if its speakers make use of electronic technology" (Crystal, 2000, p. 141).

Using information technology to assist with the survival of languages is not a new idea. In 1987 the Native Literacy Centre was established in Oaxaca, Mexico, to enable (Mexican) Indian people to read, write and publish in their own languages using computers (Russell Bernard, 1992). In 1994 Hawaiian educators established Leokī, a graphic bulletin board system installed on computers throughout immersion schools in Hawai‘i. Entirely in Hawaiian, the system facilitated email, chat, conferences, announcements and provided online dictionaries and newspapers (Warschauer & Donaghy, 1997).

The potential of information technology was also realised early by technologists working with te reo Māori. A bilingual computer system designed to support the teaching and learning of te reo Māori was created by the Kōhanga Reo movement in 1986 (Laws, 2001). A computer bulletin board called Te Wahapū was set up by the New Zealand Council for Educational Research in 1990. With its operating language in te reo Māori, Te Wahapū was designed to facilitate discussion on issues related to te reo Māori and Māori education, and to provide access to a database of new technical terms in Māori (Benton, 1996; P. Keegan, 1996). A word processor called Tā Kupu was designed in 1992 at the University of Waikato (Barbour, Cunningham, & Ford, 1993) in order to teach written reo Māori skills to Māori children. In 1993 staff at the University of Otago began to integrate reo Māori course material into a computerised laboratory so that reo Māori students could use computers to assist with their learning of te reo Māori (Laws, 1995).

Over the past two decades information technology has been used to such an extent that it has prompted the editors of a recent book of collected articles, entitled *Information Technology and Indigenous Peoples*, to state "...the use of information technology within indigenous communities and by indigenous peoples is no longer an issue of debate but a proven fact. It is no longer a dream of the future but a reality of today" (Dyson, Hendicks, & Grant, 2006, p.314).

The Internet, part of the spectrum of information technology, has the potential to make the most impact on indigenous languages. However, Grenoble & Whaley (2006) have suggested that local communities in many parts of the world are still a long way from having internet access. A report based on a survey undertaken in 2000 stated that only 35% of Māori respondents had ever used the Internet (Te Puni Kōkiri, 2001a). It appears that four years later this figure had more than doubled; a report released by Nielsen NetRatings stated that 72% of all Māori had access to the Internet (Nielsen NetRatings, 2005).

There has been considerable speculation that the Internet, and one particular component of the Internet, the World Wide Web (Web), can assist in language revitalisation. In an article titled *Can the Web Help Save My Language* Buszard-Welcher (2001) examined endangered language web sites of Native North Americans. She discovered 50 endangered language web sites with varying amounts of language material available. While stating that few of these sites had substantial amounts of language content, she did suggest that we are only beginning to realise the potential of the Web for language maintenance and revitalisation. Warschauer (2000) studied a number of Hawaiian language initiatives using the Internet and suggested that interacting in the Hawaiian language gave the students (of Hawaiian) an opportunity to explore and strengthen their sense of individual and collective Hawaiian identity. He suggested the most important role of the Internet is not its impact on transmission but its impact on will. He quoted Nancy Hornberger (Warschauer, 2000, p. 13), “language revitalisation is not about bringing a language back, it’s about bringing it forward!” Warschauer goes on to say that people will fight to maintain their language when it is not only an important part of their grandparents’ past but also of their own future. Cunliffe (2004) suggested that the presence of minority languages in the medium of the Web will become as important to language survival as having a presence in traditional media.

By having a web interface to a digital library, the digital library becomes available to all people who are able to connect to the Web. Consequently the potential and promise of the Web to assist in language revitalisation is also applicable to indigenous language digital libraries with a web interface. One such indigenous

language digital library is the Niupepa digital library, which is introduced in the following sub-section.

1.2.3 The Niupepa Digital Library

In 1996 the Alexander Turnbull Library^{xiv} made available a collection on microfiche called 'Māori Newspapers', or 'Niupepa 1842–1933'. The collection had been sourced by scouring libraries of Aotearoa for historic newspapers that were either published in te reo Māori or for a Māori audience. The collection comprised 407 microfich, covering 35 titles and 17,700 facsimiles of newspaper pages. Approximately 54% of the newspapers were written only in te reo Māori, 44% were written in parallel reo Māori and English texts and the remaining 2% were written solely in English.

The Niupepa collection contains a large amount of historical information and opinions embodying Aotearoa's early encounter history. The collection is also a rich source of an earlier reo Māori, providing a much needed text source for scholars, teachers and students of te reo Māori.

Having this collection available as microfiche was a breakthrough for researchers as it meant they could now access all the newspapers in the Niupepa collection from the one source. However, searching the collection was tedious and time consuming. The Computer Science Department at the University of Waikato resolved to redress this situation by digitising the texts, undertaking OCR (Optical Character Recognition) to allow full text searching and making the collection available online using digital library software^{xv} with a web interface.

In 2001 the Niupepa collection was made available online using the New Zealand Digital Library's (NZDL) Greenstone software^{xvi} at www.nzdl.org/niupepa.

^{xiv} An entity within the National Library of New Zealand specialising in heritage documents

^{xv} For a comprehensive explanation regarding the process of delivering the Niupepa on the Web, see T.T. Keegan et al (2001) and Apperley et al (2002).

^{xvi} For an explanation of Greenstone software see Witten & Bainbridge (2002).

The Niupepa digital library is designed with multilingual functionality with a switchable interface where the default language of the web page is normally set to reo Māori. A button on the home page allows the default language of the interface to be switched between English and reo Māori. The digital library provides three facilities for accessing the newspapers: full text search, browse by individual title and browse by date. For a full description of the Niupepa web site interface, including screenshots, see §3.1.1.

There are five reasons why the Niupepa digital library is an ideal choice to undertake usage analysis of an indigenous language. First, because the contents of the Niupepa Collection are important and valuable to Māori people it is likely that Māori people would seek to access the digital library without coercion. As the Niupepa digital library is readily available on the Web, at no charge, the likelihood of use is increased. Second, as a large proportion of the Niupepa Collection's content is in te reo Māori it is also likely to be accessed by users who are literate in te reo Māori. Third, as the interface was often set to te reo Māori it is likely that discretionary use of te reo Māori could occur. Fourth, having the ability to easily switch the interface language between the indigenous and non-indigenous language provides access to the collection by users who have no or little proficiency in te reo Māori.

The fifth reason was ease of access. The digital library hosting the Niupepa Collection was administered and maintained by the Computer Science Department at the University of Waikato, the same department where this research was undertaken. Access to the web logs and the digital library configuration files was therefore easily obtained.

1.3 Thesis Approach

Use of an indigenous language was charted by a study undertaken on the Niupepa digital library. The study was conducted from the start of January 2005 to the end of December in that year. It involved collecting all requests made to the Niupepa web site and analysing them to determine how clients used the digital library in

the indigenous language (te reo Māori). The study revealed three types of clients with significantly different usage characteristics. There were clients who preferred to use the Niupepa digital library with the interface language set to te reo Māori, clients who preferred to use the digital library with the interface language set to English and clients who preferred to use both te reo Māori and English interface language settings in their information retrieval activities.

Analysis was concentrated on the activities of te reo Māori clients. Results showed that use of te reo Māori increased if the default language of the web site was set to te reo Māori. Further analysis was undertaken to determine if the usage of the indigenous language was restricted to the locality of that indigenous population. This analysis indicated that usage of the Niupepa digital library in te reo Māori was not restricted to clients from Aotearoa; a significant number of te reo Māori international clients were using the Internet to connect and use the Niupepa digital library.

To ensure that the results were not peculiar to te reo Māori, a similar study was conducted on another indigenous language digital library. For the year of 2005 all requests to the Hawaiian Nūpepa^{xvii} digital library^{xviii} were collected and analysed as described above. These results also showed different usage characteristics amongst clients who used the interface in different languages. The results indicated that the indigenous language was used more in the Hawaiian Nūpepa digital library than it was in the Māori Niupepa digital library.

Usability studies were undertaken to determine why indigenous languages were used by clients of digital libraries and what characteristics of this environment encouraged indigenous language usage. The usability studies were undertaken in two forms—a web questionnaire and client interviews. The web questionnaire was placed on the Māori Niupepa web site for 250 days in 2006. It received 60 responses, mostly from clients who preferred to use the indigenous language (te

^{xvii} The terms Niupepa and Nūpepa are both Polynesian words for newspaper. Niupepa is the term commonly used for the Māori newspaper collection whereas Nūpepa is the term commonly used for the Hawaiian newspaper collection.

^{xviii} A full description of the Hawaiian Nūpepa digital library is given in §6.1.

reo Māori). Sixteen interviews were conducted with clients who used the digital libraries in te reo Māori, in Hawaiian and in English. The interviews involved a usage observation section and a question/answer section. The user feedback was analysed to reveal that clients used the indigenous language because of the content of the digital library, the default language setting of the digital library and because of a personal desire to use the indigenous language as often as possible.

1.4 Thesis Outline

This thesis is arranged into eight chapters. The first chapter claims that digital libraries can be used to assist in the usage, and subsequent revitalisation, of indigenous languages. The relevance of this research has been highlighted by describing the current position of te reo Māori, the importance of information technology to the revitalisation of te reo Māori and the ways in which the Niupepa digital library is suited to indigenous language usage analysis.

Transactional Log Analysis (TLA) is the method used in this research to give quantitative accounts of *how* indigenous languages are used to access information by clients of digital libraries. Chapter 2 provides a comprehensive explanation of how TLA is used to uncover reo Māori usage in the Niupepa collection. The methodology formulated in Chapter 2 is then applied in the next four chapters to determine reo Māori usage by clients of the Niupepa collection (Chapter 3), the difference between national and international reo Māori usage (Chapter 4), how the default language setting affects reo Māori usage (Chapter 5), and Hawaiian language usage by clients of the Hawaiian Nūpepa digital library (Chapter 6).

Chapter 7 uses feedback from clients to determine *why* clients prefer to use indigenous languages in bilingual digital libraries. Feedback was obtained from use of a web questionnaire and client interviews. Although the sample sizes were too small to make general claims, the questionnaire and interviews suggest that clients use the indigenous language because it matches the content, it is the default

language of the digital library and they have a desire to use the indigenous language.

Chapter 8 summarises the key findings, raises some interface design issues and asserts that digital libraries can provide an environment to support the use of indigenous languages.

2 Transaction Log Analysis

This thesis claims that digital libraries can provide an environment to support the use of indigenous languages. This claim is substantiated by defining *how* indigenous languages are used by clients of digital libraries. Transaction Log Analysis (TLA) is the method used to generate quantitative data on indigenous language usage in a digital library environment.

This chapter begins by introducing TLA and explaining some methodological decisions made in this research. The TLA undertaken on the Niupepa digital library is then described. The process involves collecting raw datasets, preparing the datasets, defining user sessions, and then analysing the sessions to determine if and how indigenous language is being used in the digital library. This TLA process is described in a simple but comprehensive manner so that the methods used can be repeated and re-applied to other bilingual or multilingual information systems.

2.1 Introduction

A transaction log is a detailed record (log) of all interactions (transactions) between a system and users (clients) of that system. This record is usually stored as a file on the system and analysis of this file is termed Transaction Log Analysis (TLA). TLA has been described as "the study of electronic recorded interactions between on-line information retrieval systems and the persons who search for information found in those

systems" (Peters (1993) cited in Jansen (2006, p. 408)). TLA can be performed by commercially available web log analysis tools; such tools were considered too inflexible and the results too limited for this particular research. Data mining techniques (also known as web usage mining or web log mining) can also be used to identify usage patterns from web logs. Data mining techniques were also considered too limited for the wide-ranging results required for this analysis of indigenous language users interacting with a digital library.

Many publications have used TLA to analyse user behaviours.ⁱ For example, Blečić et al (1998) employed TLA to reveal that many experienced difficulty with basic searching techniques when using an OPAC (Online Public Access Catalogue) system. The OPAC system's introductory screens were simplified and standardised, and key word searching was given more prominence. A second TLA determined that users' searching behaviours had improved significantly as a result of the system changes. Jansen et al (2000) used TLA to analyse queries from the Excite search engine. They reported that web users were uncomfortable with advanced methods of searching, did not use Boolean operators and did not frequently browse the results beyond the first page or so.

TLA was used by Jones et al (2000) to investigate user activity on the digital library of Computer Science Technical Reports Collection. A general description of computer science researchers' behaviours in a digital library revealed that sessions were usually very short, with few queries and the queries themselves were very simple. Chau et al (2005) used TLA to investigate whether search behaviour on a general purpose search engine, such as Excite or AltaVista, was different from search behaviour on a search engine restricted to one particular web site. They found that, in general, web users behaved similarly in each case, although there were differences in the number of queries per session, the query terms used and the query topics searched.

ⁱ A comprehensive review of transaction log analysis studies is beyond the scope of this thesis. However, Peters (1993) provides an early review and Jansen & Spink (2005) a more recent review.

TLA was used by Silverstein et al (1999) to analyse the query log of the AltaVista Search engine. The authors reported that actual web users differed significantly from the users assumed in standard information retrieval literature. Queries were short, seldom modified and the users mostly looked at the first 10 results only. They did note, however, that their methodology did not exclude data generated by non-human activity, i.e. by web robots. A technique for determining activity generated by web robots was proposed by Tan & Kumar (2002). They suggested that patterns of activity could be used to create robot classifiers that would accurately identify robot activity in transaction logs. For the TLA undertaken in this thesis, web robot activity was detected primarily by examining the User Agent and IP Addresses of the clients (see §2.3.2.2), with transaction activity also examined for unusual behaviour that could be attributed to web robots.

Jansen (2006) suggests importing transaction log data into a relational database to undertake analysis, a method popular with other researchers (Chau, Fang, & Sheng, 2005; Wang, Berry, & Yang, 2003). However, text processing scripts were used in this research to classify the data into sessions and then the sessions into language types; it was decided to continue to use text processing scripts to generate the descriptive statistics.

The TLA undertaken on the Niupepa digital library in this thesis is significant from two perspectives. First, a large amount of usage information was generated. Over 40 distinct categories of data were examined, representing significantly more categories than previous TLA studies. Second, this TLA focuses on analysing how indigenous languages are used by clients of digital libraries. Apart from preliminary analysis undertaken on the Niupepa collection by the author (Keegan & Cunningham, 2005a, 2005b; Keegan et al, 2006; Nichols et al, 2005), the TLA methodology has not otherwise been used for indigenous language usage analysis.

2.2 Collecting Datasets

The first step when undertaking Transaction Log Analysis is to collect the transaction log data. This data was collected from three sources because there were three web servers making the Māori Niupepa digital library and the Hawaiian Nūpepa digital library available to the Web in 2005.

A web server named Puka, located in Hamilton, Aotearoa,ⁱⁱ served all Aotearoa requests to the `www.nzdl.org` domain. Another web server, named Thor, located in Lethbridge, Canada, contained a mirror of the NZDL web site and served all international requests to the `www.nzdl.org` domain. The third web server, Ulukau, located in Hilo, Hawai‘i, served all requests to the `www.nupepa.org` domain. All three web servers were running the open source Apache HTTP Server software and all three were using the Greenstone Digital Library software to log similar web request information.

Each discrete request submitted by a client’s browser to one of the web servers was recorded in the web server’s log file as a single transaction. These transactions were previously termed *hits* but now are more commonly known as *requests*. The transaction log data was collected by these servers into large text files containing a defined set of details about every request received by the web servers.

The requests captured by the three servers mentioned above were copied into six web log datasets depending on certain criteria. *Table 2-2* gives a summary of the six datasets. The criteria which define these datasets are described in the following sub-sections.

ⁱⁱ Also known as Aotearoa/New Zealand and New Zealand.

2.2.1 Dataset 1 NZDL 2005

Dataset 1 consisted of all web browser requests to the `www.nzdl.org` domain name during 2005. It was gathered by combining all the requests recorded in the Puka web log with all the requests recorded in the Thor web log. In total there were 1,133,543 requests which, as a text file, required 1.16 gigabytes of disk space. As Dataset 1 contained all requests to all digital libraries housed at the NZDL web site the Niupepa digital library data was extracted by filtering out all the non-Niupepa requests. This activity is described in §2.3.2. A comprehensive analysis of Dataset 1 was undertaken to determine how reo Māori was used by clients of the Niupepa digital library. This analysis is detailed in Chapter 3.

2.2.2 Dataset 2 National and Dataset 3 International

Datasets 2 and 3 were collected so that a comparative analysis could be undertaken on the Niupepa digital library. Reo Māori usage from national (Aotearoa) users was compared with reo Māori usage from international (non-Aotearoa) users. This analysis was undertaken to determine if digital libraries can be used both nationally and internationally to assist in the revitalisation of indigenous languages.

As previously mentioned, the digital libraries available through the NZDL web site were served by two web servers. This occurred from 2003 to 2006 and involved Puka, located in the Computer Science Department at the University of Waikato, Aotearoa, and Thor, located at the Department of Mathematics and Computer Science at the University of Lethbridge, Alberta, Canada.

Requests to `www.nzdl.org` were originally received by the web server Puka. Lines of code in the Apache configuration file located on Puka pre-processed all requests to the NZDL web site. If the IP Address of the requestor did not have

.nz as a first level domain name then the request was redirected to be handled by the web server Thor. Consequently, all the requests from national users, as indicated by the user's IP Address, were served by Puka, the web server located in Aotearoa. All other requests were served by Thor, the web server located in Canada.

Creating a national (requests from Aotearoa) dataset and an international (requests from outside of Aotearoa) dataset was straightforward. The Puka web server was serving requests known to be originating from within Aotearoa so the 2005 web log from this server became Dataset 2 National Requests. The Thor web server was serving all other requests so the 2005 web log from this server became Dataset 3 International Requests.

Dataset 2 National 2005 consisted of 291,140 requests in a 268 megabyte text file. Dataset 3 International 2005 consisted of 842,403 requests stored in a 890 megabyte text file. A comparative analysis of Dataset 2 and Dataset 3 was undertaken to identify national and international differences in how te reo Māori was used by clients of the Niupepa digital library. This comparative analysis is described in Chapter 4.

2.2.3 Dataset 4 Default Reo Māori and Dataset 5 Default English

Previous TLA studies have indicated that clients do not like to alter the default settings on a digital library interface (Jones et al, 2000). The 2006 national census (Statistics New Zealand, 2006) also indicated that almost all (98.5%) of the Māori population that could speak a language, could speak English. Consequently, one could hypothesise that if the interface to the Niupepa digital library was set to English the reo Māori literate bilingual clients would use the interface in English. Alternatively, if the interface was set to reo Māori they would use the interface in te reo Māori. Dataset 4 and Dataset 5 were collected to test this theory and, thus,

determine how setting the default language of the interface influenced the usage of te reo Māori by clients of the Niupepa digital library.

The default language setting of the Niupepa interface is set by altering the CGI (Common Gateway Interface) language argument on the Niupepa default home page link. When the language argument on the default home page link is set to reo Māori, `l=mi`, as in the following example;

```
www.nzdl.org/cgi-bin/library?a=p&p=about&c=niupepa&l=mi&nw=utf8
```

then all the texts, commands and buttons on the user interface of the Niupepa web site are displayed in te reo Māori. When the language argument on the default home page link is set to English, `l=en`, as in the following example;

```
www.nzdl.org/cgi-bin/library?a=p&p=about&c=niupepa&l=en&nw=utf8
```

then all the texts, commands and buttons on the user interface of the Niupepa web site are displayed in English. The default home page link is stored in a macro file called *nzdlhome.dm* which is located on the web servers. This link is used when web browsing software requests the default Niupepa home page.

If a client subsequently alters the interface language setting of the Niupepa web site, the altered language will stay in effect until the client reloads the home page. The home page URL does not include any CGI arguments; thus, a reload on this page will cause the predefined default settings, as in the examples listed above, to be reinstated. If a client bookmarks a page of the NZDL and the bookmark includes the CGI `L` Argument, then a request for the bookmarked page will not invoke the default language setting of the Niupepa interface.

At the start of 2005 the default language of the Niupepa interface was set to reo Māori. Throughout 2005 this default language was altered on a monthly basis. This frequency was chosen to cause as little disruption to the user as possible, yet seasonal usage activity would still be captured in both datasets. In the months of January, March, May, July, September and November the default language argument of the Niupepa interface was set to reo Māori. In the months of February, June, August, October and December the default language argument of

the interface was set to English. The dates and times that the default language was altered are listed in *Table 2-1*.

Set on PUKA		Set on THOR		Default Lang.
1 st January	00:00 am	1 st January	00:00 am	reo Māori
4 th February	1:32 pm	3 rd February	5:35 pm	English
1 st March	9:50 am	28 th February	1:50 pm	reo Māori
1 st April	9:15 am	31 st March	2:42 pm	English
<i>Unknown</i>	<i>Unknown</i>	<i>Unknown</i>	<i>Unknown</i>	reo Māori
2 nd May	8:50 am	1 st May	2:48 pm	reo Māori
9 th June	9:49 am	8 th June	3:51 pm	English
1 st July	3:56 pm	30 th June	9:59 pm	reo Māori
1 st August	9:17 am	31 st July	3:19 pm	English
1 st September	8:54 am	31 st August	2:56 pm	reo Māori
3 rd October	10:00 am	2 nd October	3:03 pm	English
1 st November	10:45 am	31 st October	2:48 pm	reo Māori
2 nd December	9:59 am	1 st December	2:02pm	English

Table 2-1: The dates, times and (new) default languages that were set on the Niupepa web site in 2005

Some time during April the default language of the Niupepa interface was (by mistake) set to reo Māori. This setting may have occurred as a result of some software upgrades that happened at that time. Because it is not known for certain exactly when in April this default language setting was changed all April requests were excluded from Dataset 4 and Dataset 5.

The default language was set to te reo Māori for 198 days (54.2%) and to English for 137 days (37.5%). There were 30 days (8.2%) when, because of uncertainty in the default language setting, the web log data was not used.

Dataset 4 Default Reo Māori contains all requests received by the Niupepa web site when the default interface language was set to te reo Māori (as outlined in *Table 2-1*). Dataset 4 Default Māori consists of 598,260 requests and is stored in

608 megabyte file. Dataset 5 Default English contains all requests received by the Niupepa web site when the default interface language was set to English (as outlined in *Table 2-1*). Dataset 5 Default English consists of 407,145 requests and is stored in a 417 megabyte file. Both datasets contain requests from national (Aotearoa) users and international (non-Aotearoa) users.

A comparative analysis of Datasets 4 and 5 was undertaken to identify whether altering the default language of the Niupepa interface influenced the usage of te reo Māori by clients of the Niupepa digital library. This comparative analysis is detailed in Chapter 5.

2.2.4 Dataset 6 Hawaiian Nūpepa 2005

Dataset 6 Nūpepa consisted of all requests to the `www.nupepa.org` domain name in 2005. This domain is responsible for serving the Hawaiian Nūpepa digital library. This digital library is similar to the Māori Niupepa digital library in that it made available a large collection of historic newspapers written in an indigenous language using Greenstone software. However, there were some interface differences between the Hawaiian Nūpepa digital library and the Māori Niupepa digital library. These are listed in *Table 6-1*. Because of these differences, minor changes needed to be made to the scripts that generated the Transaction Log Analysis statistics for Dataset 6. For example, requests in Hawaiian were collated as opposed to requests in te reo Māori; counts were undertaken of middle sized documents as opposed to just large and preview documents; data about preference pages and help pages could not be collected as these pages were not available with the Hawaiian Nūpepa interface.

On the 3rd of December 2005 the Nūpepa web site was moved onto a new web server. The new web server had a default setting that utilised its ‘web performance cache’. This default setting had the effect of making all requests appear in the log as if they came from the IP Address 128.171.24.14; the web server’s IP address.

This situation continued until mid January when the web site administrators noticed what was happening and reset the default setting so that the originating requestors' IP addresses were correctly recorded in the web logs. While normal activity continued it had all been recorded incorrectly as being from the one IP address. The method used to define sessions does not rely on IP addresses; thus, it can be ascertained that over this period; there were 3,174 single requests, 133 exploratory session requests, and 17,384 usage sessions requests, a total of 20,691 requests.

However, there is a small chance that the incorrect recording of the IP addresses compromised the session data. If two new users made a request to the Nūpepa collection within the same second they both would have been allocated the same unique identifier, meaning both users would appear in the logs as the same entity. Such a situation would compromise the data. Although the chances of this situation occurring were very small it was decided to remove all requests from the IP address 128.171.24.14. Removing these requests had the effect of removing most of the requests recorded in December and reducing the size of the web log by 6.7%. The effects of removing these requests can be seen in *Figure 6-10*.

Dataset 6 Nūpepa consists of the 2005 web log file created by the Ulukau web server, with most of the requests recorded in December removed. In total there were 286,383 requests to the Nūpepa digital library stored in a 286 megabyte file.

A comprehensive analysis of Dataset 6 Nūpepa was undertaken to determine how the Hawaiian language was used by clients of the Hawaiian Nūpepa digital library. This analysis is detailed in Chapter 6.

2.2.5 Summary of the Six Datasets of Web Logs

The six datasets of web logs are summarised in *Table 2-2*.

Summary of Web Log Datasets					
	Description	Web Server	Timeframe (all 2005)	Requests	Size (Mbyte)
1	NZDL	Puka, Thor	Jan-Dec	1,133,543	1,160
2	National	Puka	Jan-Dec	291,140	268
3	International	Thor	Jan-Dec	842,403	890
4	Default reo Māori	Puka, Thor	Jan, Mar, May, Jul, Sep, Nov	598,260	608
5	Default English	Puka, Thor	Feb, Jun, Aug, Oct, Dec	407,145	417
6	Hawaiian Nūpepa	Ulukau	Jan-Dec	286,383	268

Table 2-2: Information about web log datasets listing descriptions, web server, when recorded, requests and dataset file size

2.3 Preparing Datasets

As mentioned earlier, undertaking Transaction Log Analysis (TLA) involves four discrete steps: collecting the datasets, preparing the datasets, defining sessions, and undertaking analysis on the sessions.

Preparing the datasets involves removing unwanted data. This reduces processing time and ensures that the results are not influenced by unrelated data. Two types of data were removed from the datasets: redundant CGI (Common Gateway Interface) arguments and extraneous requests. The resultant datasets only contained data relevant to analysing how indigenous languages are used by clients of digital libraries.

2.3.1 Removing Redundant CGI Arguments

Each time a request is received by a web server, details about that request are recorded into a log file. The request details that are recorded vary depending on the type of web server and how the digital library has been configured. The Greenstone digital library software is usually configured to capture a large amount of data with each request. Below is an example of the entire request data recorded for a single request, with colour added to assist with explanation:

```
/cgi-bin/library 218-101-90-11.dialup.clear.net.nz [Thu
Jun 09 10:09:25 +1200 2005] (a=p, b=0, b1=0, b2=0,
bclaboutdesc=, bclcfgchanged=0, bclclone=0,
bclclonechanged=0, bclclonecol=, bclcontactemail=,
bcldirname=, bcldelete=0, bcleconf=0, bclesrce=0,
bclfromsrce=0, bclfullname=, bclinfochanged=0,
bclinput=, bclinputnum=3, bclinputtype=, bcltmp=, bcp=,
beu=, bft=, bnu=, bp=, bt=0, c=niupepa, cc=, ccp=0,
ccs=0, cfgfile=, cl=, cm=, cq2=, ct=0, d=, de=, debc=0,
dm=, ds=, dsbc=0, e=, el=prompt, er=, f=0, fc=1, fqa=0,
fqc=, fqf=, fqk=, fqn=4, fqs=, fqv=, g=, gc=0, gt=0,
h=, h2=, hd=0, hl=1, hp=, hs=0, ifl=, il=1, j=, j2=,
k=1, ky=, l=mi, m=50, n=, n2=, nl=, nw=utf-8, o=20,
p=about, pc=, pfd=0, pfe=0, pfl=0, pld=10, ple=10,
pll=10, ppnum=0, pptext=, pw=, pxml=0, q=, q2=, qb=0,
qf=0, qt=0, qto=3, r=1, rd=0, s=0, st=1, t=1, u=0, ua=,
uan=, ug=, uma=listusers, umc=, umnpw1=, umnpw2=,
umpw=, umug=, umun=, umus=, un=, us=invalid, v=0,
w=utf-8, x=0, xx=0, z=218.101.86.158-1118228138)
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
```

This set of data consists of the following sections:

`/cgi-bin/library` - the directory on the web server where all the CGI (Common Gateway Interface) programs are stored. The CGI programs

allow data to be passed between host serving software and the client browsing software.

`218-101-90-11.dialup.clear.net.nz` —the IP Address and/or host name of the client browser that submitted the request.

`[Thu Jun 09 10:09:25 +1200 2005]` —the time stamp indicating when the request was received (Greenwich Mean Time is used).

```
(a=p, b=0, b1=0, b2=0, bclaboutdesc=, bclcfgchanged=0,
bclclone=0, bclclonechanged=0, bclclonocol=,
bclcontactemail=, bcldirname=, bcldelete=0,
bcleconf=0, bclesrce=0, bclfromsrce=0, bclfullname=,
bclinfochanged=0, bclinput=, bclinputnum=3,
bclinputtype=, bcltmp=, bcp=, beu=, bft=, bnu=, bp=,
bt=0, c=niupepa, cc=, ccp=0, ccs=0, cfgfile=, cl=, cm=,
cq2=, ct=0, d=, de=, debc=0, dm=, ds=, dsbc=0, e=,
el=prompt, er=, f=0, fc=1, fqa=0, fqc=, fqf=, fqk=,
fqn=4, fqs=, fqv=, g=, gc=0, gt=0, h=, h2=, hd=0, hl=1,
hp=, hs=0, ifl=, il=1, j=, j2=, k=1, ky=, l=mi, m=50,
n=, n2=, nl=, nw=utf-8, o=20, p=about, pc=, pfd=0,
pfe=0, pfl=0, pld=10, ple=10, pll=10, ppnum=0, pptext=,
pw=, pxml=0, q=, q2=, qb=0, qf=0, qt=0, qto=3, r=1,
rd=0, s=0, st=1, t=1, u=0, ua=, uan=, ug=,
uma=listusers, umc=, umnpw1=, umnpw2=, umpw=, umug=,
umun=, umus=, un=, us=invalid, v=0, w=utf-8, x=0, xx=0,
z=218.101.86.158-1118228138)
```

—the CGI arguments; settings used by the CGI programs to assist with information transfer. Many of these arguments are not used.

`"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"`

—the name and compatibility of the client's web browser (this is called the user agent string).

The CGI arguments from the web logs that are used and are relevant to this research are briefly described as:

<code>a=p</code>	action; p=page, q=query, d=document
<code>c=niupepa</code>	collection; which collection the request is associated with
<code>d=</code>	document identification (only set if a=d)
<code>g=</code>	document type (text, preview image, full image)
<code>k=1</code>	case-matching on queries (1=no matching, 2=case matching)
<code>l=mi</code>	language setting of interface (mi=reo Māori, en=English)
<code>m=50</code>	maximum results for query
<code>p=about</code>	page type (used if a=p); help=help page, about=about page
<code>q=</code>	query string; query term(s) submitted in a query
<code>s=0</code>	stemming (0=stemming off, 1=stemming on)
<code>t=1</code>	search type; some (t=1) or all (t=0)
<code>w=utf-8</code>	character set encoding
<code>z=218.101.86.158-1118228138</code>	unique user identifier (see §2.4.1)

A considerable amount of data is stored for each request in the log file. Almost three quarters of this data is redundant or not relevant to this particular research. Consequently, the first step in the preparation of the web logs was to remove all the CGI arguments that were not relevant to the analysis on indigenous language research. A text processing script, written in the Perl programming language, a Perl script, was created to remove the redundant data recorded for each request (see §2.5 for a summary of Perl scripts). This Perl script, called *process_raw_logs.pl*, was run on the web log datasets. Using the request data example listed above, the requests data remaining after *process_raw_logs.pl* was run is:

```
/cgi-bin/library 218-101-90-11.dialup.clear.net.nz [Thu
Jun 09 10:09:25 +1200 2005] (a=p, b=0, cl=, d=, gg=,
gp=, h=, j=, k=1, l=mi, m=50, p=about, q=, r=1, s=0,
t=1, v=0, w=utf-8, z=218.101.86.158-1118228138)
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
```

The request data example is thus reduced from 1041 characters to 262 characters, a reduction of 75%. Reducing each individual request data reduces the overall size of the web log datasets by a similar ratio, reducing the storage and processing requirements without compromising any of the relevant data.

2.3.2 Removing Extraneous Requests

The second stage of preparing the datasets is the removal of extraneous requests. The extraneous requests include: requests to other (non-Niupepa) digital libraries located in the NZDL, requests from non-human activity, requests made where the interface language is not reo Māori or English, requests with incomplete or incorrect settings and local requests submitted by developers of the Niupepa digital library. More than 80% of the requests collected in Dataset 1 were classified as extraneous requests and were removed from this indigenous language usage analysis. A description of these removals is given below.

Removing Non-Niupepa requests

The web log files generated by the Greenstone software record requests made to all digital library collections available at the New Zealand Digital Library (NZDL) web site (www.nzdl.org). There were 67 individual digital libraries available at the NZDL web site in 2005. The only requests appropriate to this research are requests to the Niupepa Collection (www.nzdl.org/niupepa). The Perl script *process_raw_logs.pl* was modified to also remove requests from Dataset 1 that were not requests to the Niupepa web site (i.e. requests were removed that did not have the CGI argument setting of `c=niupepa`).

Dataset 1 initially comprised 1,133,543 total raw requests. However 929,178 (82%) were requests to non-Niupepa digital libraries and were subsequently removed from this dataset. This left 204,365 requests remaining in Dataset 1 available for further processing.

Removing Requests From Web Robots

This research is concerned with indigenous language usage of the digital libraries by real users (i.e. human activity). Steps were undertaken to remove all requests that could be identified as being requested from Web robots. A web robot is a program that automatically travels the Web accessing web pages, and links on those pages.ⁱⁱⁱ Web robots are also known as web crawlers, web spiders and web wanderers. Two methods were used to identify requests from web robots: filtering for previously identified web robots, and analysing access behaviour for unusual, non-human like activity.

Previously identified web robots are defined by user agent strings and IP Addresses.

One task of the Perl script *process_raw_logs.pl* was to check the user agent strings stored in the web logs against a list of previously identified web robot user agent strings. The list was obtained from www.pgts.com.au/downloads.data in a file called *robots_list.txt*. At the time of the analysis, this file listed 327 robot user agent strings. When the Perl script *process_raw_logs.pl* found a user agent string match with a listed robot user agent string it removed the matched request from the dataset.

A similar method was used to check IP Addresses for known robot IP Addresses. The *process_raw_logs.pl* script also included code to check the IP Addresses in the dataset against a list of previously identified robot IP Addresses. This list was obtained from www.iplists.com and at the time of the analysis, contained 4,426 IP Addresses that had been previously identified as undertaking web robot activity. When the Perl script *process_raw_logs.pl* found an IP Address match with a listed robot IP Address it removed the corresponding request from the dataset.

ⁱⁱⁱ For more information about web robots see www.robotstxt.org/wc/robots.html and for lists of known web robots see www.dnsstuff.com.

These methods of identifying web robots were applied to Dataset 1 and identified a total of 292 web robots submitting 16,081 requests to the Niupepa web site.

Analysing access behaviour to identify web robots often occurs at a later stage in the TLA process. Tan and Kumar (2002) suggest that this method can be used in isolation to identify web robots. However, for the purposes of this research it was used as a secondary methodology. Identifying web robots by access behaviour involved defining the requests into sessions, defining the sessions into session types and then examining the session types over set time frames.

In the single request sessions (for a definition of single request sessions see §2.4.2) it became quite obvious that there were four IP Addresses with an abnormally high count of unique identifiers in a two month time period towards the end of 2005. These IP Addresses were creating 150–300 unique identifiers per month when the average number of unique identifiers created for that session type was just two per month. These four IP Address were investigated on the web site www.dnsstuff.com and were shown to be web robots that had recently been created. They were added to the list of known web robots and then the script *process_raw_logs.pl* was rerun.

Subsequently the Perl script *process_raw_logs.pl* identified a total of 296 web robots submitting 17,177 requests to the Niupepa Log. These requests were removed from the dataset.

While every endeavour has been made to remove web robot requests from the dataset it can be argued that requests from web robots may still exist in the log file. Perhaps some of the requests are from web robots that have not yet been identified. However, these requests are believed to be small in number and not likely to affect the overall results. Also, it is unlikely that such web robots requests submit queries and access actual pages of the Niupepa Collection, which is the focus of this TLA into how indigenous languages are used by clients of digital libraries.

Removing Requests With Inappropriate Language Settings

The analysis for this research was only concerned with requests submitted when the interface language was set to either reo Māori or English. All other requests were discarded. Aside from these two settings, the client has the ability to set the language of the Niupepa interface to one of 38 other languages. In addition, if the client accessed the site by using bookmarks, and some of the CGI arguments were altered throughout the year with software upgrades, as happened on a number of occasions, then the values in some of the CGI arguments may have been stored incorrectly. A typical example of this is a null value or a 00 value being stored in the CGI `l Argument`. The script *process_raw_logs.pl* also included code to remove all requests where the `l Argument` was not set to either `mi` or `en`. For Dataset 1 there were a total of 1018 requests (0.09%) removed because of an inappropriate `l Argument` setting.

Removing Requests With Other Incorrect Settings

Aside from the removal of requests outlined in the previous paragraphs, the Perl script *process_raw_logs.pl* was also configured to remove requests from the dataset where other CGI arguments were stored incorrectly. A request was removed from the dataset if there was an incorrect setting in the CGI `c (collection) Argument` or an incorrect setting in the `IP Address Argument`. The IP Addresses were also examined to detect the IP Addresses of computers that belonged to staff involved with maintaining and upgrading the digital library or web site. If the IP Address of a request was identified as belonging to a Niupepa staff member's computer, the request was removed from the dataset. The activity under investigation was the use of the Niupepa digital library by clients, not support staff.

2.3.3 Extraneous Requests Summary

A list of the extraneous requests removed from the initial raw log file Dataset 1 is displayed in *Table 2-3*. After the unwanted requests were removed, the remaining requests became the prepared log file, Dataset 1 Niupepa 2005. This completes the second step of TLA, preparing the datasets. The prepared log file was then available for the third step of TLA, defining sessions.

	Requests Removed	Requests
Initial Log file (Dataset 1)		1,133,543
Non-Niupepa Collection Requests	929,178	
Web Robot Requests	17,177	
Inappropriate language Argument	1,018	
Incorrect Collection Argument	4	
Incorrect IP Address	457	
Local IP Address	708	
Prepared Log File (Dataset 1 Niupepa)		185,001

Table 2-3: *Number of requests removed from initial raw log file Dataset 1 to create the prepared log file.*

2.4 Defining Sessions

The next stage in the TLA process was to collate the prepared datasets of requests into sessions so that indigenous language usage could be examined. Collating the requests into sessions involves three procedures: defining unique identifiers, defining session types and defining language of session types. These three procedures are explained in the next three sub-sections.

2.4.1 Defining Unique Identifiers

In order to collate requests into sessions the requests must be associated with an individual client. Identifying individual clients raises a whole range of issues. It is virtually impossible from a web log to identify specific individuals as insufficient client-explicit data is stored. Also, if specific clients could be identified then their permission would need to be sought before any personal information could be retained about them. The solution is not to identify clients as specific individuals but rather to identify them as unique identities whose personal details remain unknown. To identify clients as unique identities, a unique identifier is created and recorded with all activity undertaken by that client.

Two common techniques are currently used to associate clients with unique identifiers. The first technique is to use as a unique identifier the IP Address of the computer submitting the request. The second technique involves creating a unique identifier for each client and storing it on the client's computer in the form of an HTTP cookie^{iv}. Both techniques were available for the TLA undertaken on the Niupepa digital library as both techniques are possible using the Greenstone software.

As described in §2.3.1, the Greenstone software records the IP Address of the client browser with each request in the web logs. This IP Address can be used as a unique client identifier.

The Greenstone server software also uses HTTP cookies (known simply as cookies) by first querying the client's browser for a unique identifier cookie stored from the Niupepa web server. If the client's browser is able to respond with the Niupepa unique identifier cookie then the value stored in this cookie is assigned to the `Z Argument` for all hits generated by that client/browser (for an example of the `Z Argument` see §2.3.1 above). If the client doesn't have a Niupepa unique identifier cookie, one is generated by concatenating the client's IP Address with

^{iv} An HTTP cookie is a small text file created by a web server and saved by a client's web browser on the client's computer. The text of the cookie is sent to the web server each time the client makes a request to that web server.

the current UNIX time stamp. This method ensures no two unique identifiers will be assigned the same value. From then on this value, stored by client's browser as a Niupepa unique identifier cookie in the client's personal account and assigned by the web serving software to the `Z Argument`, can be used to uniquely identify the client.

There are advantages and disadvantages to both the IP Address technique and the cookie/`Z Argument` technique of defining unique identifiers.

The IP Address is always readily available; it is always supplied with requests and hence, using an IP Address to identify users is the method currently favoured by most web log analysis software. However, while the IP Address is often computer specific, it is not user specific; consequently if multiple users share one computer in sequence then several quite distinct sessions by different users might be classified as one single session. Further, people connecting through a shared dialup connection may also have a common IP Address and again, multiple users' activities might be classified as one single session.

Using cookies and the `Z Argument` to define unique identifiers eliminates the issue of users connecting through a shared dialup and it eliminates the issue of users using the same computer, providing they are logged onto the computer as different users. However, this technique will not identify separate users when multiple users are logging onto a shared computer using the same username (for example, if two or more people use a public library computer to access the collection). Also, users who do not enable cookies will never appear in the session data because a new unique identifier will be generated for each new request. Users who clear their stored cookies regularly will have a new unique identifier generated each time the cookies are cleared and this action will again affect the conclusions drawn in the session analysis.

IP Address or Cookies/`Z Argument` to Define Unique Identifiers?

Which technique was used to define unique identifiers in TLA undertaken in this thesis? The decision was straightforward. It was decided to use the Cookies/`Z`

Argument technique as although there were 4% fewer requests defined as actual usage sessions, there was more confidence that multiple users were not being defined as a unique identity.

2.4.2 Defining Session Types

Collating a prepared set of numerous requests into sessions of activity by individual clients is the core element of TLA that allows usage analysis. The most common method for collating requests into sessions is a timeout method. Using this method, sessions are defined as a sequence of two or more requests originating from a single client where the time between requests is no longer than a set interval. This set interval is usually 30 minutes.

The timeout method was used in this research to define sessions; however, the time interval was set higher than normal at 60 minutes. The time interval was increased because Niupepa users might be expected to spend extended periods reading the large amount of text on some of the newspaper documents and because the gap between requests might be exaggerated by web caching (which could mask intermediate requests).

It was decided to separate all prepared requests into one of three types of sessions: single request sessions, exploratory sessions and usage sessions. This decision was made for two reasons. First, as well as allowing for actual usage activities to be examined, it also allowed for the examination of exploratory activities, giving insights into why separate user groups may explore the web site but then not go on to access content stored in the web site. The second reason for defining the requests into three session types was to add another filtering stage so that the requests that were eventually defined as usage session requests were as unequivocal as possible.

Single request sessions were defined as requests submitted by a client when there was no other request made by that same client within the 60 minute time interval. These requests are not identifiable sessions as such but rather a random smattering of requests by seemingly different clients. Single requests could occur because of a client seeing one page of the web site and then deciding not to view any more pages. Single requests could also occur when a client has disabled cookies on their web browser.

An exploratory session was defined as a sequence of requests submitted by a single client with no more than a 60 minute time period between requests, no queries were submitted and no collection documents were accessed. Exploratory sessions constituted clients who accessed interface home pages, the help pages and/or the preferences pages but did not request any actual newspaper documents.

A usage session was defined as a sequence of requests submitted by a single client with no more than a 60 minute time period between requests, and, queries are submitted and/or collection documents requested. Usage sessions represent definitive activity by clients who were actually using the digital library to access content.

2.4.3 Defining Session Language Types

In order to analyse how indigenous languages were used by clients of digital libraries sessions were further defined into session language types. Session types were defined into two groups—those sessions where the indigenous language was used to undertake the activity and those where the non-indigenous language was used. For Dataset 1 Niupepa 2005, the sessions were divided into those sessions where the interface language was set to te reo Māori and those sessions where the interface language was set to English.

Defining the session language types for single request sessions is straightforward. Each request has the `L Argument` set to `mi` or `en`, that is, the language of the interface set to reo Māori or English. These interface language settings can be easily counted and compared. However, with exploratory sessions and usage sessions it is not so easy to classify the session as being a reo Māori session or an English language session as the sessions often comprise a combination of both language settings. Some decisions were made about what percentage of requests in te reo Māori constituted defining a session as a reo Māori session and what percentage of requests in English constituted defining a session as an English language session.

Exploratory sessions were assigned to session language types depending on which language the exploratory session was *mostly* conducted in. The `L Argument` was counted for each request in each session. If the `L Argument` was set to `mi` for more than 50% of the requests in the session, then that session was classified as an exploratory session *mostly* conducted in the medium of te reo Māori. For simplicity, the phrase ‘exploratory session mostly conducted in the medium of te reo Māori’ has been abbreviated to an ‘exploratory reo Māori’ session (`expMi`). Conversely, if the `L Argument` was set to `en` for more than 50% of the requests in the session, then that session was classified as an exploratory session *mostly* conducted in the medium of English. This phase has been abbreviated to an ‘exploratory English’ session (`expEn`). On those occasions when the `L argument` was set to one language for half of the session and the other language for half of the session then the session was classified by the language setting that the final request of the session was set to.

For usage sessions, the language classification was a little more complicated. Earlier research (Keegan & Cunningham, 2005a) had indicated that there were in fact three language classifications for usage sessions of the Niupepa digital library: usage sessions mostly in te reo Māori (abbreviated to ‘usage reo Māori’ (`usgMi`)), usage sessions mostly in the English language (abbreviated to ‘usage English’ (`usgEn`)), and bilingual sessions, that is, sessions where the user interface was set to reo Māori and English for a significant amount of the session (abbreviated to ‘usage Bilingual’ (`usgBi`)). After much consideration it was felt

that the earlier research may have classified many short sessions as bilingual sessions when, in fact, they were most likely monolingual clients taking more than one request to switch the interface language. Consequently, a new set of criteria for defining session language types was created, the details of which follow.

A usage session was defined as **usgMi** if it met the following criteria:

- L Argument = mi in more than 50% of the requests
- L Argument = mi in exactly 50% of the requests and
L Argument = mi in final request
- Session not defined as usgBi session

A usage session was defined as **usgEn** if it met the following criteria:

- L Argument = en in more than 50% of the requests
- L Argument = en in exactly 50% of the requests and
L Argument = en in final request
- Session not defined as usgBi session

A session was classified as **usgBi** if it met *all* of the following criteria:

- session must contain 10 or more requests
- L Argument =mi in more than 20% of the requests
- L Argument =en in more than 20% of the requests
- must contain 3 or more switches of interface language

The final requirement of the usgBi session, that bilingual sessions must contain at least 3 switches of interface language, was used to ensure that this grouping of sessions contained users who were specifically switching the language of the interface to assist with information retrieval. It was designed to eliminate clients who logged into the web site, used the interface in one language and then tried it in the other language but, on realising that this language was not as accommodating, then reverted back to the original language setting.

2.4.4 Collating Sessions

Two Perl scripts were written to collate the requests into sessions. A Perl script called *create-sessions.pl* collated all the prepared requests into one of the three types of sessions: single request sessions, exploratory sessions and usage sessions. The sessions were first collated using IP Address as a unique identifier and then collated again using the `Cookie/z argument` as a unique identifier. Both methods of creating sessions were undertaken for comparison purposes only, sessions created using the `Cookie/z argument` method were the only sessions used in the subsequent TLA processes.

A Perl script called *split-sess-lang.pl* divided exploratory and usage sessions into language sessions. All exploratory sessions were divided into one of two language session types; exploratory reo Māori sessions (`expMi`) and exploratory English sessions (`expEn`). All usage sessions were divided into one of three language sessions types, usage reo Māori sessions (`usgMi`), usage English sessions (`usgEn`) and usage bilingual sessions (`usgBi`).

2.5 Perl Scripts and Statistical Definitions

Text processing scripts, written in the Perl programming language were used to prepare the raw datasets of requests and collate these requests into appropriate session types for the analysis of indigenous language usage. Three Perl scripts have already been mentioned. The Perl script *process_raw_logs.pl* was written to remove redundant CGI arguments and extraneous requests. The Perl scripts *create-sessions.pl* and *split-sess-lang.pl* were written to define session types and session language types. Seven additional Perl scripts were written to generate a wide range of statistical data. The Perl scripts were written by Gareth Judsen under supervision in 2005/2006 while Gareth was a student of the Computer

Science Department at the University of Waikato. A summary of the ten Perl scripts is given in *Appendix A Perl Scripts Summary*.

A primary function of the Perl scripts was to generate statistical data to determine how indigenous languages are used by clients of digital libraries. This statistical data was output in text files which were transferred into spreadsheets. Spreadsheets were used to assist with comparative analysis and to generate graphs and tables. A listing of the statistical data generated by the Perl scripts is given in *Appendix A Perl Scripts Summary*.

The Perl Scripts were also used to generate indigenous language statistics on apparent reading and language switching by clients of a digital library. Apparent reading and language switching are explained in the next two sub-sections, followed by a sub-section explaining how statistical significance is defined.

2.5.1 Apparent Reading

The apparent read time is the length of time between two requests in a session. It is coined an apparent read time because it is impossible to confirm from the web logs that the user only spent their time reading that particular page or document before requesting the next page or document. With the large number of requests for the web interface pages and the Niupepa collection documents, and as we are only using these times to make comparisons against other times calculated in a similar manner, it is proposed that apparent read times can provide useful insights into indigenous language activity in a digital library.

There are two types of apparent read times—apparent page reads and apparent document reads. An apparent page read is the length of time between when a client requests a page of the web interface and when the client selects some other page or document. Apparent page reads are the lengths of time a client spends interacting with the web site interface. An apparent document read is the length of time between when a client requests a collection document, for example a newspaper page, and when the client selects some other page or document.

Apparent documents reads are the lengths of time that users spend reading the material of the collection.

Statistical results of apparent read times are discussed with other TLA statistical results in Chapters 3–6.

2.5.2 Language Switching Data

A language switch is defined as the changing of the interface language from reo Māori to English or from English to reo Māori. For the purpose of identifying the point at which the language switching was occurring, the sessions were divided into equal tenths. The point at which the language switch occurred was associated with the closest $1/10^{\text{th}}$ position in the session. The Perl script called *split-sess-lang.pl* was modified to generate language switching data. The language switching data included the number of language switches for each usage session type and the position of language switch within the usage session. Statistical results of interface language switching are discussed with other TLA statistical results in Chapters 3–6.

2.5.3 Determining Statistical Significance

Statistical significance is used in this research to decide whether differences observed between indigenous and non-indigenous users were real or whether the differences could have been caused by chance. Differences are observed in a wide range of areas and then mathematics is used to determine the probability that these differences are not caused by chance. If the probability (of chance) is low, then it can be assumed that the differences are real and are caused by some other factors. If the probability (of chance) is high then it can be assumed that there is statistically no significance in the difference.

The probability factor is referred to as a p value. Mathematical probabilities, like p values, range from 0 to 1, where 0 indicates there is no probability (that the difference is because of chance) and 1 indicates certainty (that the difference is because of chance). Generally speaking, if the p value is calculated to be less than 0.05 then the difference is said to be statistically significant. However, to give even more certainty, a p value lower than 0.01 will be used in this thesis to indicate differences are statistically significant.

The chi square test can also be used to determine the significant difference in data. It is particularly applicable for use in results from TLA as it can be easily applied to test for significance in bivariate tables. The statistics produced from the TLA in this research is often presented in bivariate tables. Consequently, p values, generated from chi square tests, are used to determine if the differences in these values are statistically significant.

2.6 Summary

This chapter has explained in detail the Transaction Log Analysis methodology used to define how indigenous languages are used by clients of digital libraries. The collection of the six datasets of web log requests used in this research has been described. The methods used to prepare those requests and filter out redundant CGI arguments and extraneous requests have been described. The definitions used to collate the resulting requests into session types and language of session types have also been explained. The Perl scripts used to generate statistical information from the language session types have also been summarised. The next stage is to examine the statistical data generated by the TLA process to determine how indigenous languages were used by clients of digital libraries.

Statistical data generated about indigenous language usage of the Niupepa Collection is analysed in Chapter 3. Statistical data analysing the difference between national and international indigenous language usage of the Niupepa

collection is revealed in Chapter 4. Statistical data interpreting the influence of the default language setting on indigenous language usage is examined in Chapter 5. Statistical data generated about indigenous language usage of the Hawaiian Nūpepa Collection is analysed in Chapter 6. These four chapters comprehensively examine a very large amount of data on the issue of how indigenous languages were used by clients of digital libraries.

3 Reo Māori Usage of the Niupepa Digital Library

Chapter 3 discusses how te reo Māori was used by clients of the Niupepa digital library in 2005.

A prime question that must be asked before the discussions begin is, was te reo Māori used by Niupepa clients? Results from Transaction Log Analysis (TLA) indicated that te reo Māori was used significantly by clients in single request sessions, exploratory sessions and usage sessions. Having established that te reo Māori was used, the next stage is to examine how it was used by Niupepa clients.

Data from exploratory sessions gave insights into how clients explored the Niupepa interface in te reo Māori but then did not go on to use the Niupepa digital library. This data showed that Niupepa exploratory sessions in te reo Māori (expMi) were short, with the Niupepa home page being the interface page that received the most requests.

Data from the usage sessions was then examined to determine how reo Māori clients used the Niupepa interface, how reo Māori clients accessed documents and how reo Māori clients submitted queries. These three topics give a comprehensive account of how te reo Māori was used by Niupepa clients.

Statistics generated by TLA were also used to compare reo Māori/English language usage and examine bilingual use. The reo Māori/English language comparison reveals different session lengths, different reading times and different querying behaviours between usgMi clients and usgEn clients. The bilingual

usage (usgBi) analysis reveals usgBi sessions to be long, with a focus on using queries to access documents.

This chapter begins by presenting a description of the interface screens and options that are available to clients of the Niupepa digital library as context for discussing reo Māori usage.

3.1 Niupepa Interface Description

The Niupepa digital library has a multilingual interface that serves a collection of legacy newspapers written mostly in te reo Māori. The Niupepa digital library can be found at www.nzdl.org/niupepa. The web site offers different methods for accessing the newspaper documents and the documents themselves are available in various formats. As well as newspaper documents, additional information about the newspapers is available from the Niupepa web site. The methods available for accessing the Niupepa documents, the document formats and additional documents available to the Niupepa clients are displayed in the screenshots below.

The home page of the Niupepa web site is displayed in *Figure 3-1*. The interface is displayed with the default language set to reo Māori; the usual default language setting for this web site. The language of the interface can be easily changed to English by clicking on the English hyperlink which is located towards the top of the home page. The home page is the only web site page on which the language of the interface can be directly alternated between English and reo Māori. Elsewhere the interface language can be switched by using the Preferences page (as displayed in *Figure 3-2*).



Figure 3-1: The home page of the Niupepa web site (reo Māori interface).

The Niupepa home page is divided into three sections. Along the top right of the page are the Āwhina (Help), Kōwhiringa (Preferences) and NZDL buttons which link to their respective web pages. Below these buttons is the title of the page (kāinga or home) and then the access bar. The access bar contains three primary methods to access the documents. The Rapu (Search) button invokes the search engine shown in *Figure 3-3*. The Tānga Pukapuka (Series) button will display a series list of all the newspapers as shown in *Figure 3-4*. The Ngā Rā (Dates) button will show a list of newspaper documents published in a particular time period, as shown in *Figure 3-5*. The home page text gives background information about the Niupepa collection.

The Preferences (Kōwhiringa) page of the Niupepa digital library web site is displayed in *Figure 3-2*.

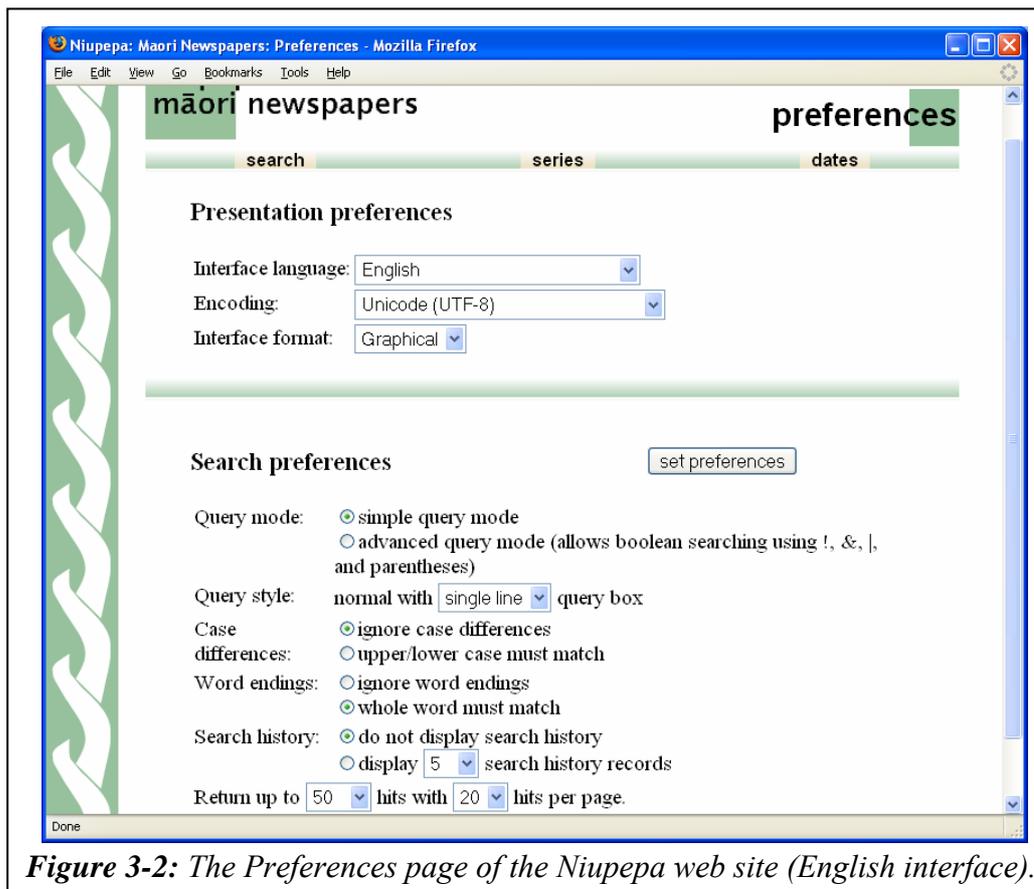


Figure 3-2: The Preferences page of the Niupepa web site (English interface).

Accessing Niupepa Documents

There are three methods by which clients may access Niupepa documents. Clients can submit a query as displayed in *Figure 3-3*. This example shows the results page following the submission of a query for ‘Paratene Ngata’ⁱ. The second method is to select a periodical from the series list, as displayed in *Figure 3-4*. Clicking a periodical title from this page will display a list of issues available from that newspaper. The third method of accessing Niupepa documents is by using the Date facility, as displayed in *Figure 3-5*. This example shows all the available documents published in the year 1900.

ⁱ Paratene Ngata is an ancestor of the author from the Ngāti Porou tribe.

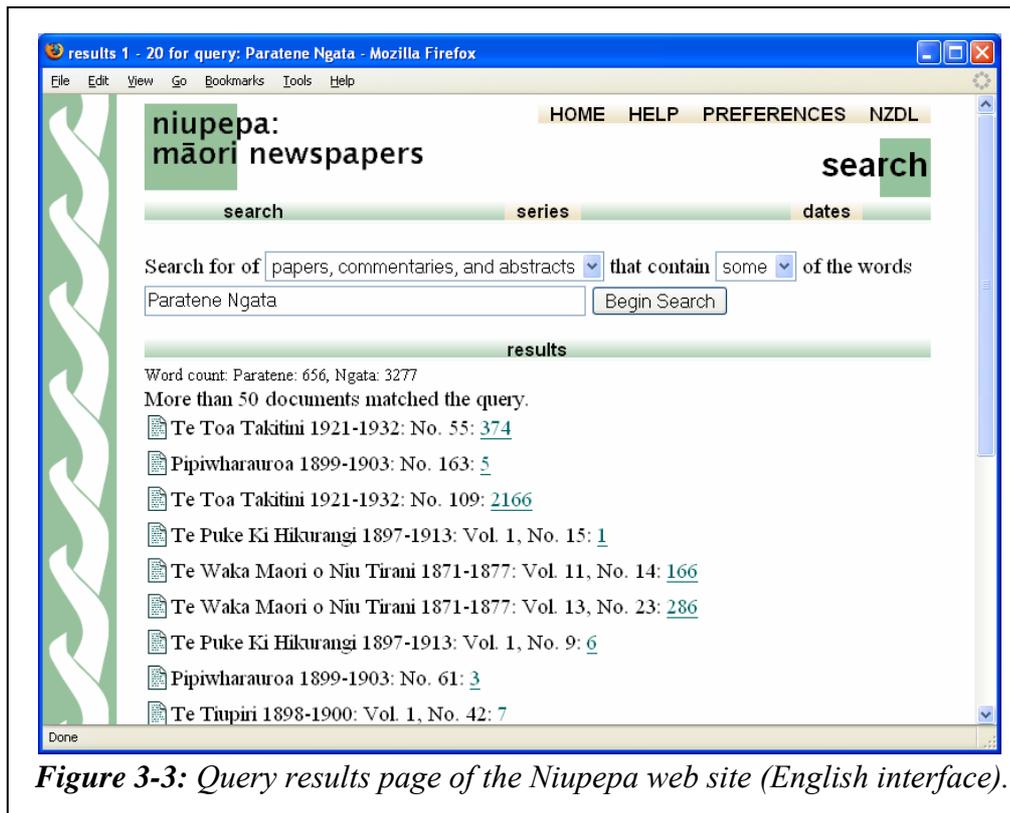


Figure 3-3: Query results page of the Niupepa web site (English interface).



Figure 3-4: Series list of the Niupepa web site (English interface).

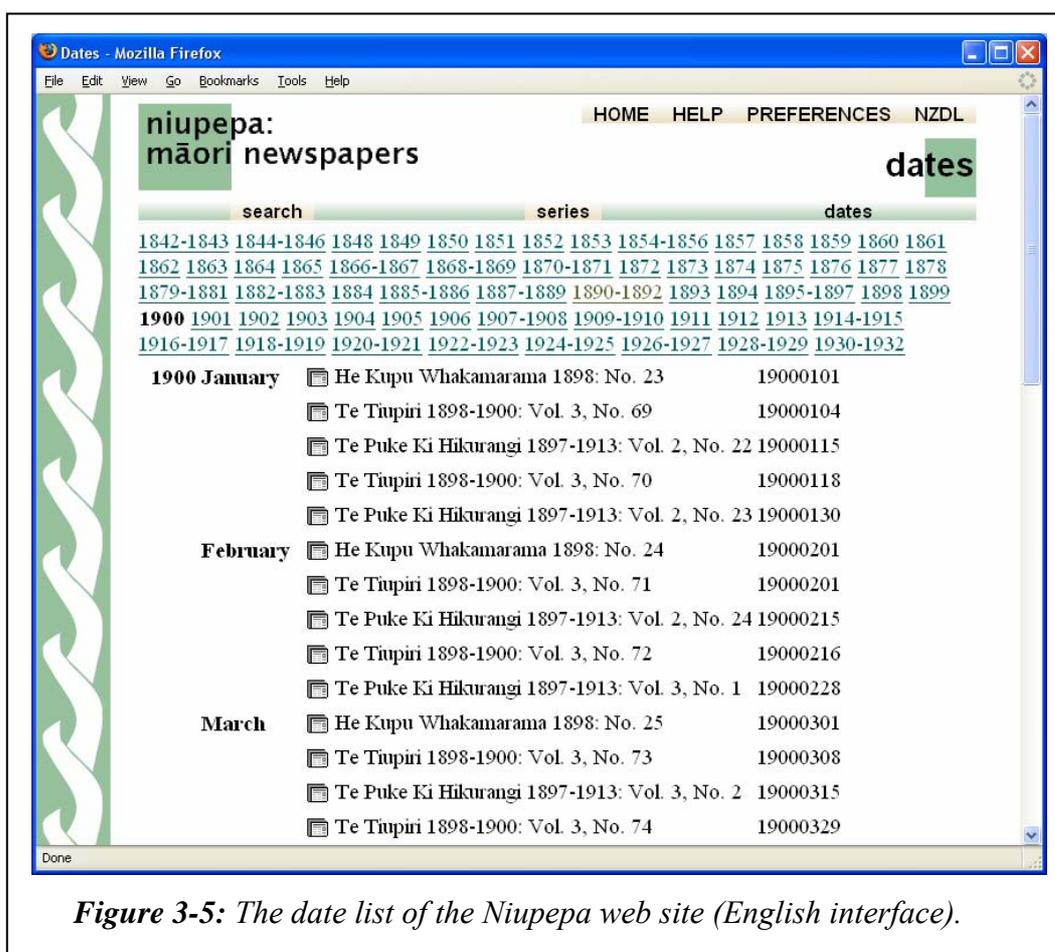


Figure 3-5: The date list of the Niupepa web site (English interface).

Figure 3-3 also shows the two query parameters that the client can alter on the Query page. The first parameter is papers, commentaries and abstracts. By default the texts of the newspapers, the bibliographic commentaries and the English abstracts are searched collectively but an option is available to search separately on any one of these resource types. The bibliographic commentaries and the English abstracts are described below. The second query parameter is the Some/All parameter which is used when more than one query term is entered. It gives the user the option to seek documents that contain *some* of the query terms or documents that contain *all* of the query terms.

Niupepa Document Formats

The Niupepa web site displays the Niupepa documents in three formats, extracted text, preview image and full size image. An example of the extracted text format is shown in *Figure 3-6*, an example of the preview image format is shown in *Figure 3-7*, and an example of the full size image format is shown in *Figure 3-8*. The examples listed show a document that has been requested following a query for the terms 'Paratene Ngata'. Highlighting of the query terms only occurs in the extracted text version of the document.

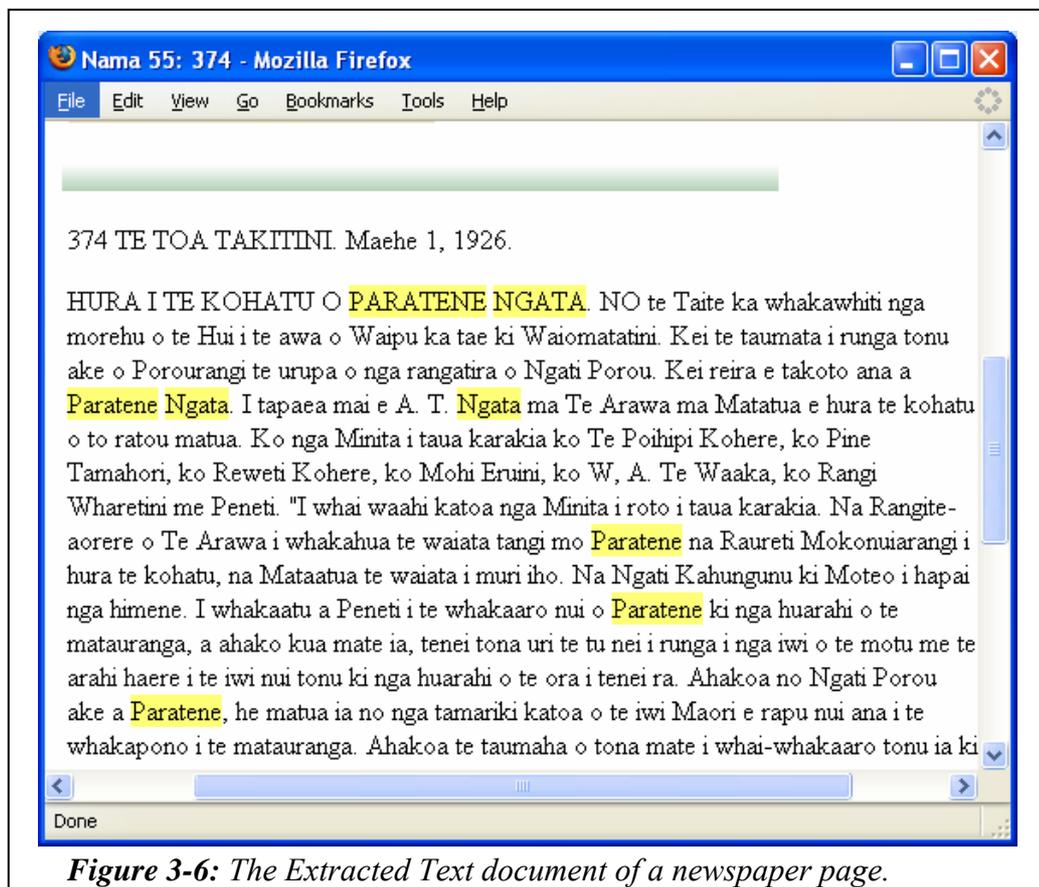


Figure 3-6: The Extracted Text document of a newspaper page.

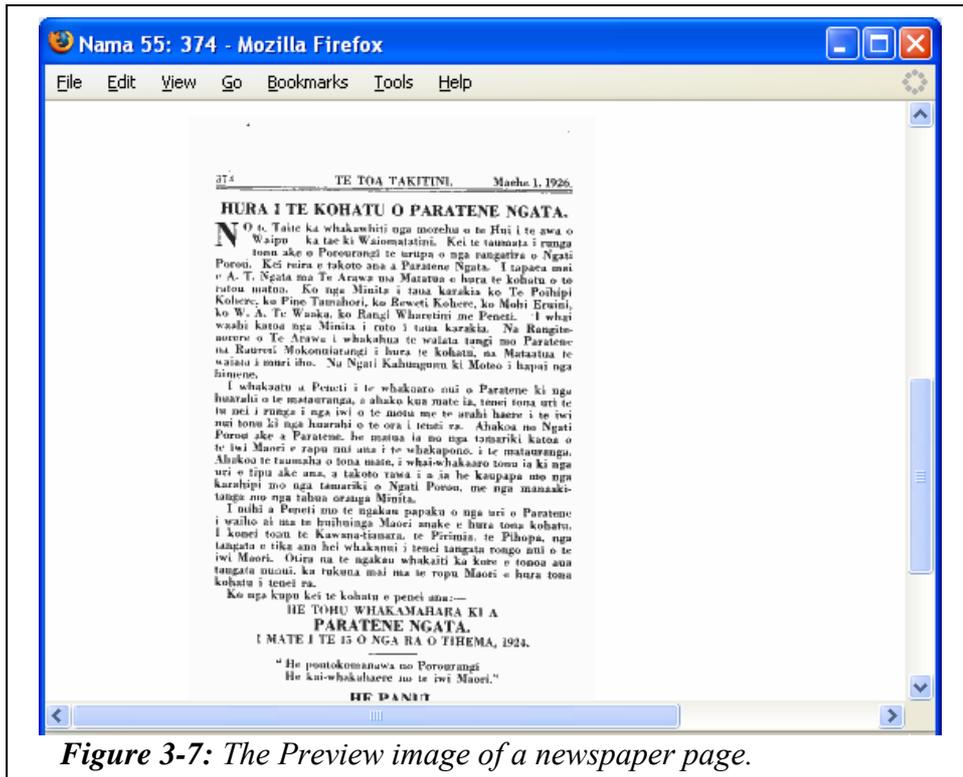


Figure 3-7: The Preview image of a newspaper page.

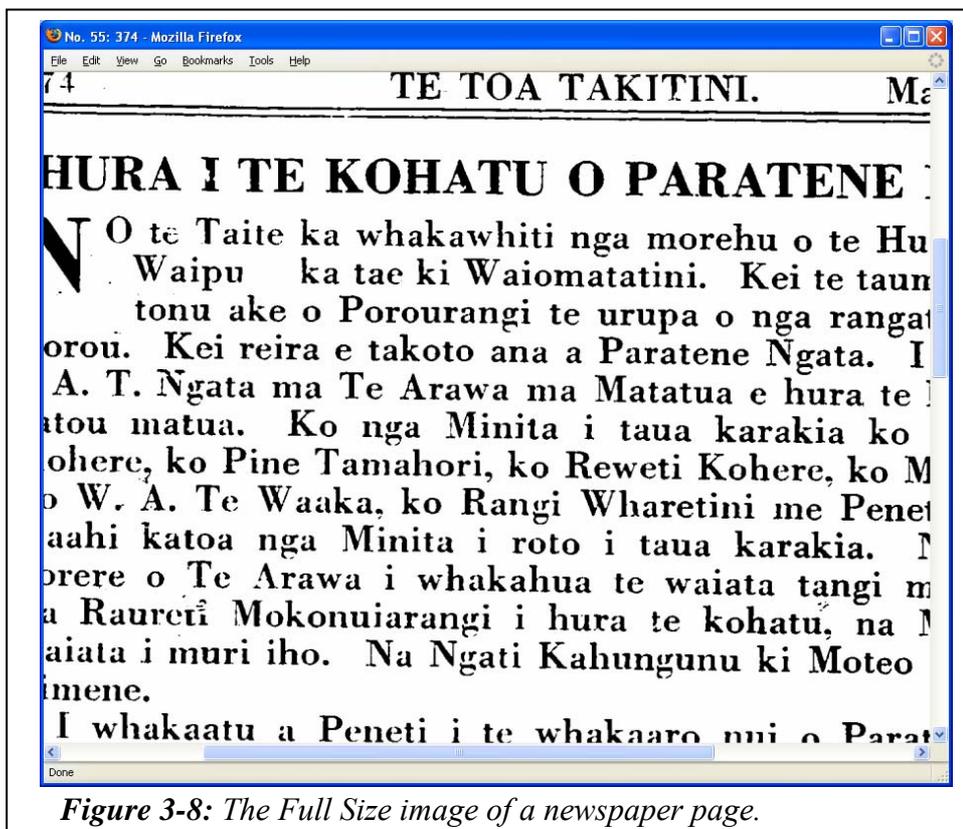


Figure 3-8: The Full Size image of a newspaper page.

The Niupepa web site also makes available two additional types of information: English abstracts and bibliographic commentaries. The English abstracts were compiled by a team of researchers at the University of Auckland headed by Dr Jane McRae. The abstracts are summaries (in English) of over 40% of the reo Māori newspapers and give non-Māori literate users an understanding of what has been written. An example of an English abstract is shown in *Figure 3-9*.

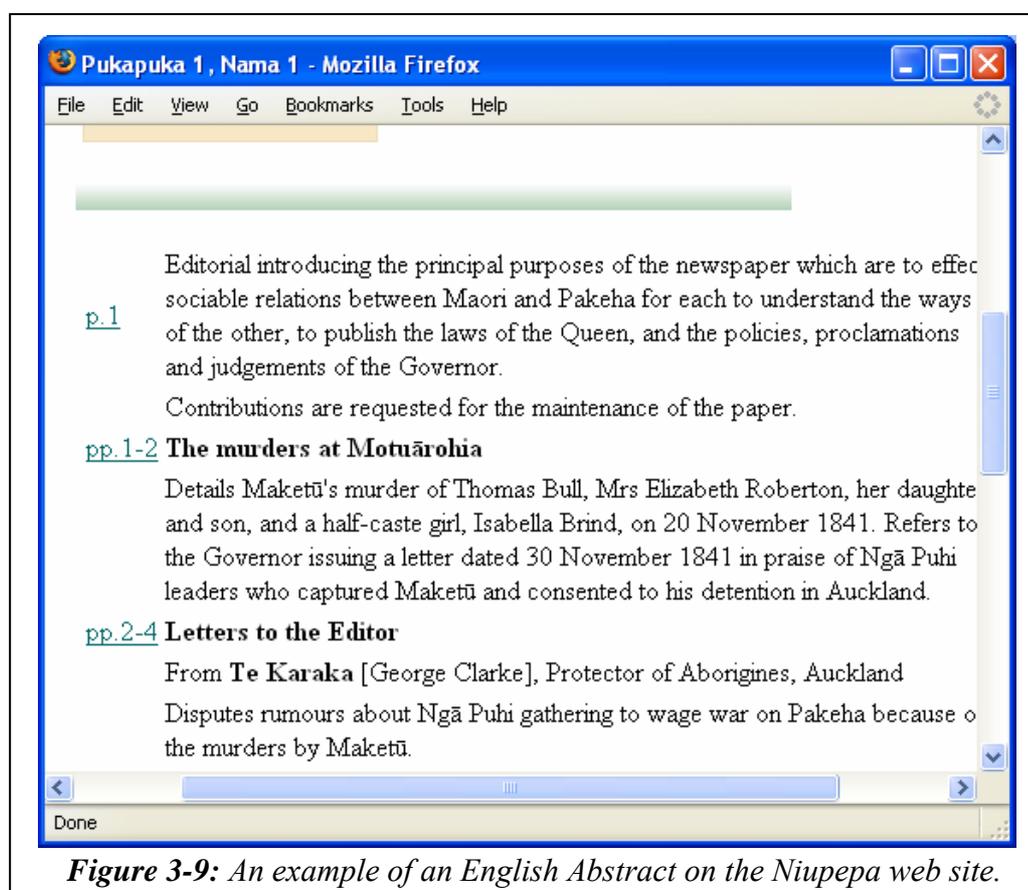


Figure 3-9: An example of an English Abstract on the Niupepa web site.

The bibliographic commentaries were written by Gail Dallimore (Dallimore, 1990) and give background information about each periodical, including publication details, subject matter and where surviving originals can be sourced. An example of a Bibliographic Commentary is shown in *Figure 3-10*.

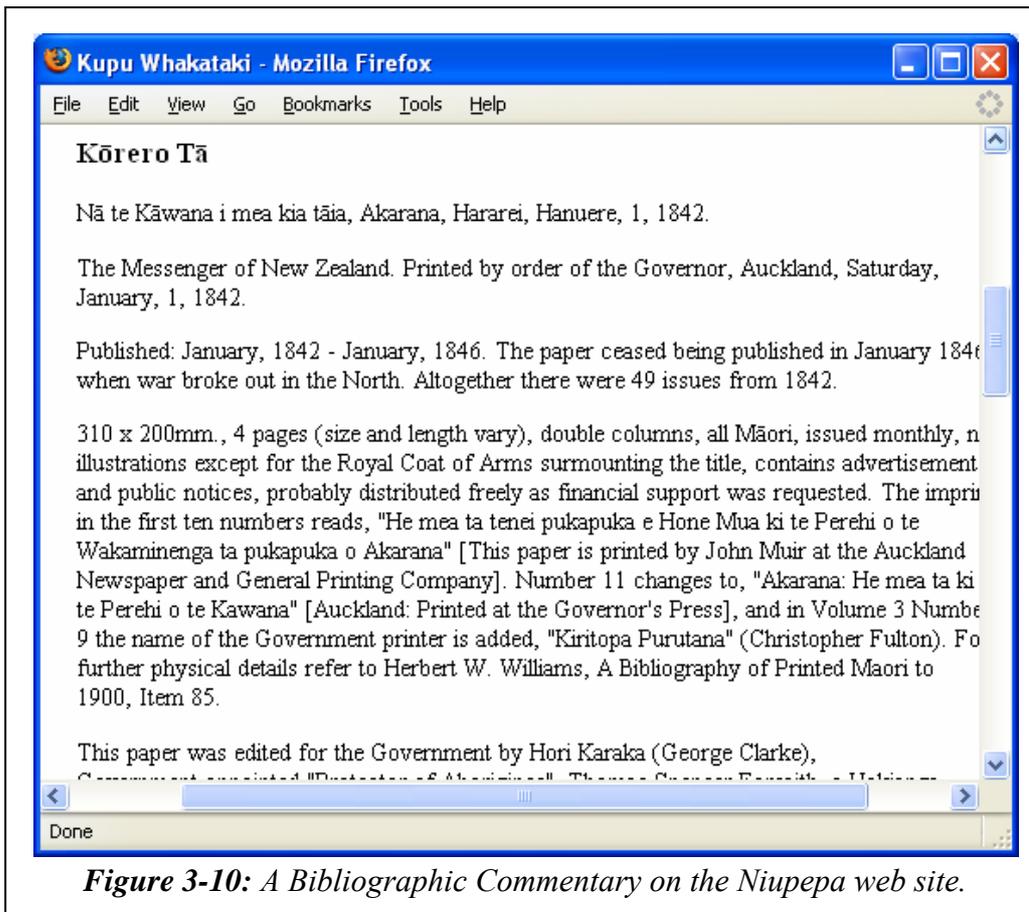


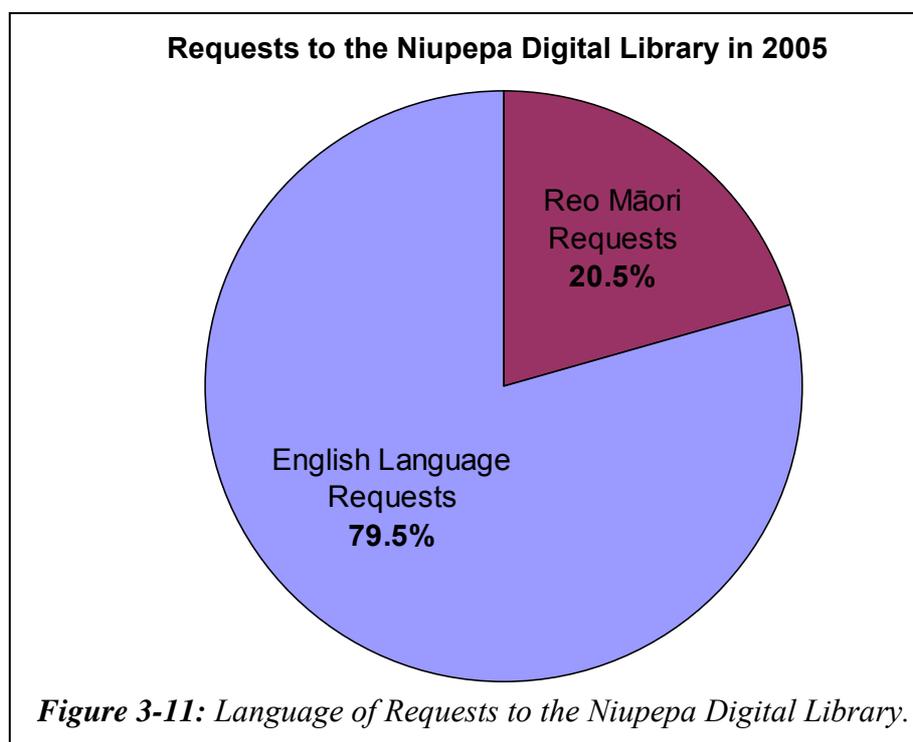
Figure 3-10: A Bibliographic Commentary on the Niupepa web site.

Figure 3-9 and *Figure 3-10* bring to a close the description of the Niupepa interface. The following section of this chapter will begin to examine how this interface was used in te reo Māori by clients of the Niupepa digital library.

3.2 Was Te Reo Māori Used by Niupepa Clients?

The first question to be answered from the TLA statistics is, was te reo Māori used by clients of the Niupepa digital library?

The Niupepa web site received 1,133,543 raw requests in 2005. The first request was logged at 12:02am January 1st 2005 and the last request for 2005 was logged at 11:58pm December 31st 2005. This raw log of requests was filtered to remove all the unwanted requests (as described in §2.3.2) leaving a filtered total of 185,001 requests (see *Table 2-3*). Of this total, 37,936 (20.5%) were requests with the interface language set to reo Māori and 147,065 (79.5 %) were requests with the interface language set to English. These requests are displayed in *Figure 3-11*.

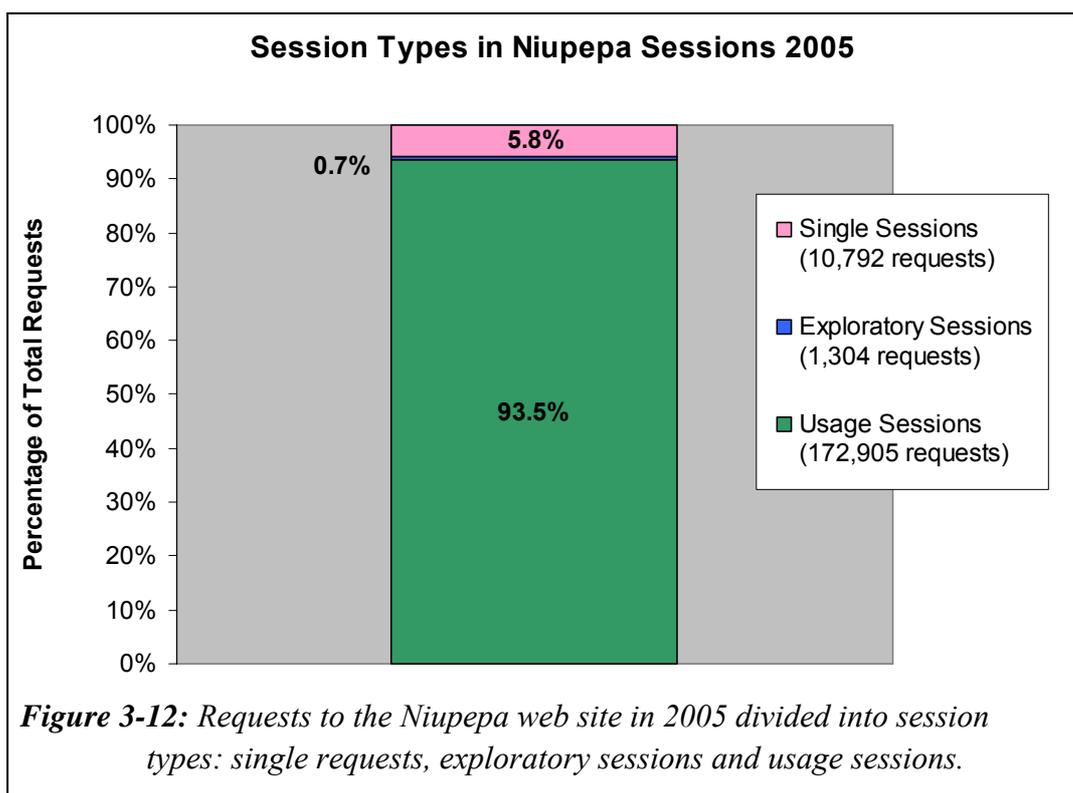


Consequently, we can state that te reo Māori was used in approximately 20% of client requests to the Niupepa digital library. However, was it used in all three of the session types?

3.2.1 Was Te Reo Māori Used in all Session Types?

The TLA process separated the requests into the three session types (as described in §2.4.2). There were 10,792 requests (5.8%) classified as single request sessions, 1,304 requests (0.7%) classified into 567 exploratory sessions, and 172,905 requests (93.5%) classified into 6,744 usage sessions. The percentages of requests in the different session types are displayed in *Figure 3-12*.

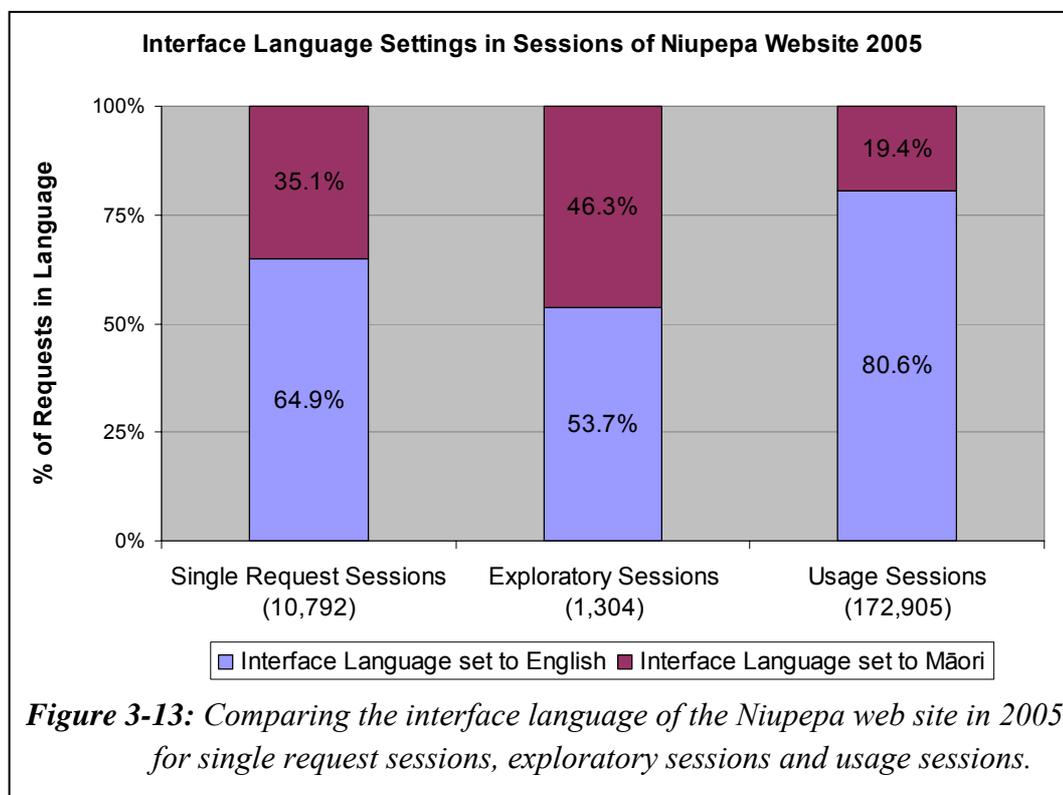
Figure 3-12 shows the majority of filtered requests were usage session requests, that is, requests from clients who were actively engaging with the Niupepa digital library to read newspaper documents and/or undertook a query to locate those documents.



The interface language settings from the TLA were examined for the three session types. For single request sessions there were 3,792 requests when the interface language was set to te reo Māori and exactly 7,000 when the interface language was set to English. For exploratory sessions, there were 604 requests when the

interface language was set to te reo Māori and 700 when the interface language was set to English. For usage sessions, there were 33,540 requests when the interface language was set to te reo Māori and 139,365 when the interface language was set to English. These figures have been calculated as percentages and are graphed in *Figure 3-13*.

Figure 3-13 clearly shows that te reo Māori was being used by clients of digital libraries in all three session types.



As the Niupepa web site's default language was set to reo Māori for approximately 60% of the time used in the web log data, one could suppose that there would be a 60/40 ratio of single requests in reo Māori and English. The 35/65 ratio suggests that some of the requests may have been incorrectly defined as single requests, perhaps caused by users disabling cookies or by network and browser caching which would prevent common requests being received by the Niupepa web server.

The Exploratory Sessions show an almost 50:50 ratio of reo Māori use, confirming that te reo Māori was being used by clients of the Niupepa digital library to explore the Niupepa collection.

The usage sessions show that te reo Māori was being used in almost 20% of the usage requests. This percentage indicates significant use and 33,540 usage requests in te reo Māori give sufficient data for further TLA to determine exactly how te reo Māori was used by clients of the Niupepa digital library.

3.2.2 When Was Te Reo Māori Used?

There are three timeframes to be considered in analysing when te reo Māori was used by clients of the Niupepa digital library. What time of the day was it used? What day of the week was it used? And what month of the year was it used?

Requests from exploratory sessions in te reo Māori (expMi) were combined with requests from usage sessions in te reo Māori (usgMi) and the percentages were plotted in *Figures 3-14, 3-15 and 3-16*. As a language comparison, the English exploratory and usage requests (expEn and usgEn) were also plotted in *Figures 3-14, 3-15 and 3-16*.

*Figure 3-14*ⁱⁱ shows the requests grouped per hour. Te reo Māori was mostly used between the hours of 9:00am and 10:00pm, with a clear peak between 1:00pm and 3:00pm. A similar usage pattern was noted when the digital library was being used in English; however, the English language requests did not peak to the same extent just after midday.

ⁱⁱ Note: this data is only taken from national requests where it is certain that the time zone of the client matched the time zone recorded in the Web log.

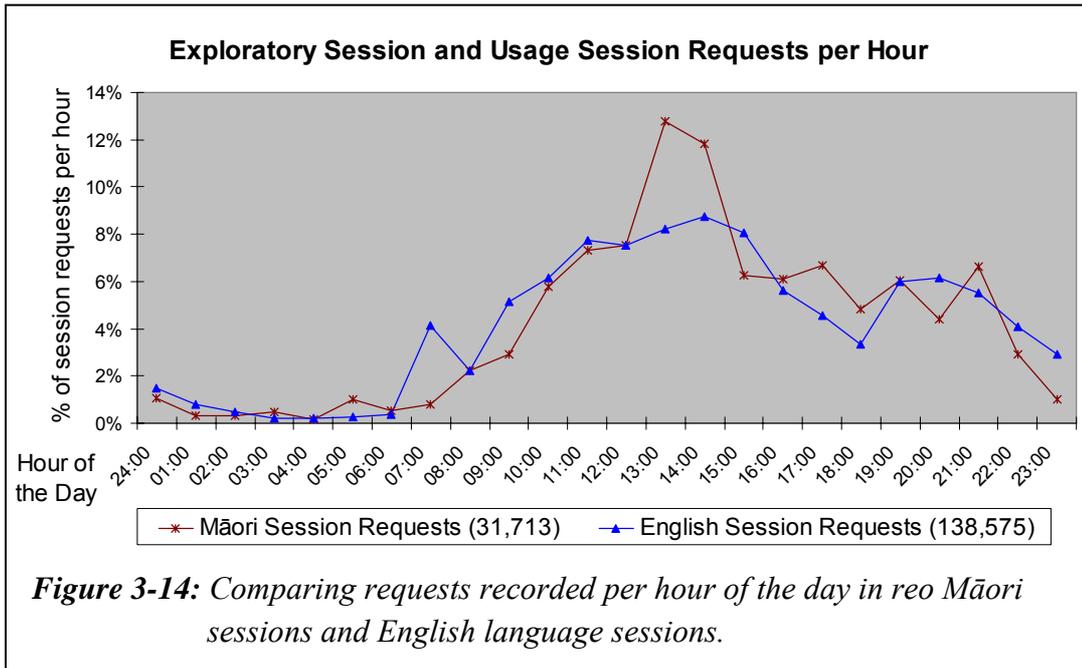
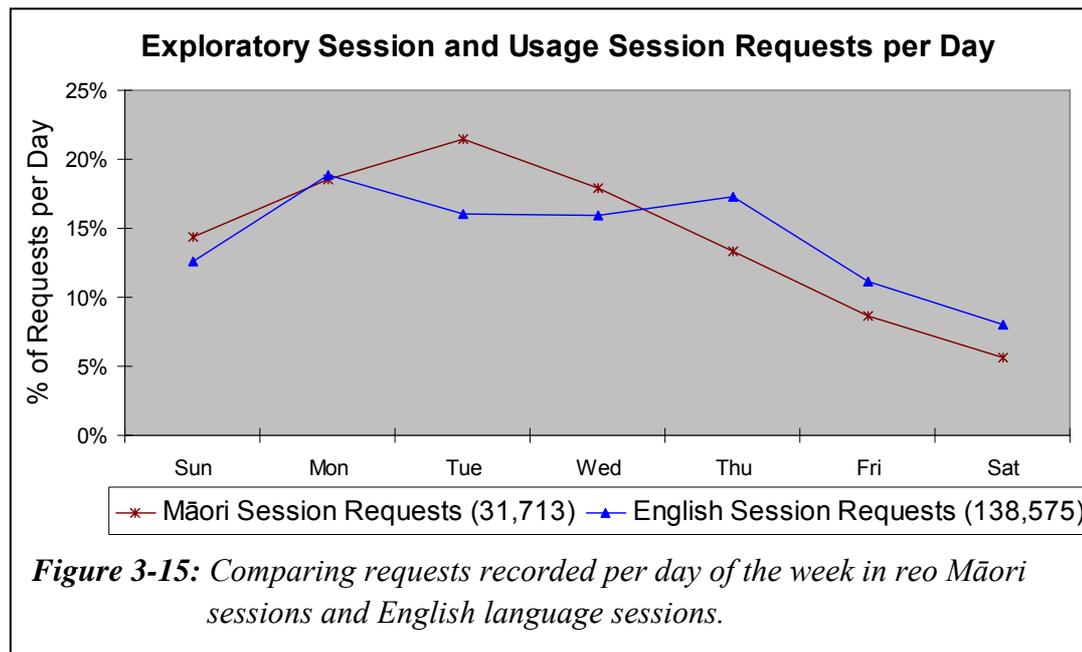
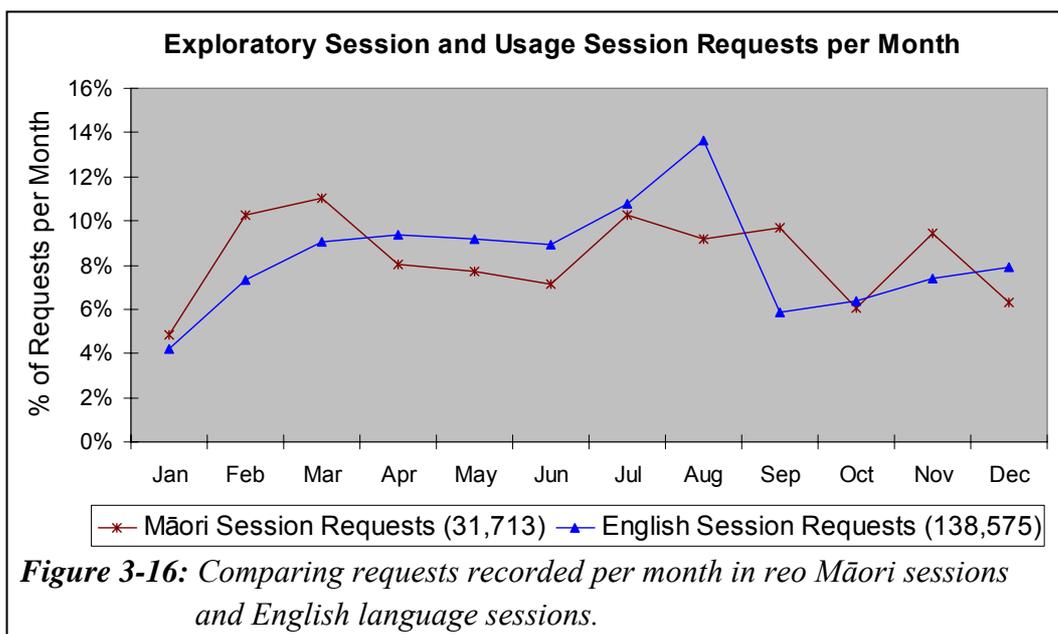


Figure 3-15ⁱⁱⁱ shows use of te reo Māori was mostly occurring on Mondays through to Wednesdays, with the least use occurring on Fridays and Saturdays. This usage was similar to the usage seen in the English language, although English language clients showed a preference to use the digital library on Mondays through to Thursdays.



ⁱⁱⁱ See previous note.

Figure 3-16 shows requests in te reo Māori grouped by month, as they occurred in 2005. Excluding January, the variance in months is never more than 3.2% indicating that te reo Māori was used consistently throughout the year. This consistency was not reflected in the English language usage, which showed a peak in August, followed by a trough in September.



In summary, te reo Māori was used by clients of the Niupepa digital library in 2005, mostly between 9:00am and 10:00pm with a peak between 1:00pm and 3:00pm, mostly on Mondays, Tuesdays and Wednesdays, and reasonably consistently throughout the year.

Now that it has been established that te reo Māori was being used and when it was being used, the next stage is to identify exactly how it was being used. The remainder of this chapter describes client behaviour and is the result of a comprehensive Transaction Log Analysis undertaken on the Niupepa web logs.

3.3 Exploratory Behaviours

Requests made by clients that explore the Niupepa digital library (but then do not go on to use it) represent less than 1% of the total filtered requests. While this is a low number of requests, it does represent 5.7% of the combined number of exploratory and usage sessions. This data can highlight how these types of clients, exploratory users, use the Niupepa interface before deciding not to access any documents or submit any requests. In particular, this section will examine exploratory sessions by users who explore in the medium of te reo Māori (expMi).

There were 1304 requests recorded in 567 exploratory sessions for the Niupepa collection in 2005. A high proportion, 84.0%, of the sessions were less than three minutes and a high proportion, 93.7%, consisted of three requests or less. Note that two requests was the minimum requirement for classification as a session.

The question we now ask is, how was te reo Māori used in exploratory sessions by clients of digital libraries? To answer this question the exploratory sessions were subdivided into the two language categories. There were 180 exploratory sessions where the Niupepa web site interface language was mostly set to reo Māori (expMi). In these sessions the interface language was set to reo Māori for 81.2% of the requests. There were 387 exploratory sessions where the Niupepa web site interface language was mostly set to English (expEn). In these sessions the interface language was set to English for 70.6% of the requests.

TLA was used to examine three characteristics of exploratory sessions in te reo Māori: the length of expMi sessions, the apparent reading times in expMi sessions and the interface pages requested in expMi sessions.

3.3.1 How Long Were Exploratory Sessions?

Session length is measured in two categories—duration in time (seconds) and the number of requests. The comparative lengths of expMi and expEn sessions in seconds are displayed in *Table 3-1*. The comparative lengths of expMi and expEn sessions by number of requests are displayed in *Table 3-2*.

expMi			expEn		
Seconds	Sessions	%	Seconds	Sessions	%
Total	180	100.0%	Total	387	100.0%
0–1	36	20.0%	0–1	38	9.8%
2–9	17	9.4%	2–9	85	22.0%
10–29	40	22.2%	10–29	116	30.0%
30–179	50	27.8%	30–179	94	24.3%
180+	37	20.6%	180+	54	14.0%

Table 3-1: Comparisons of length (in seconds) of expMi and expEn sessions.

expMi			expEn		
Requests	Sessions	%	Requests	Sessions	%
Total	180	100.0%	Total	387	100.0%
2	133	73.9%	2	329	85.0%
3	34	18.9%	3	35	9.0%
4	9	5.0%	4	11	2.8%
5–6	4	2.2%	5–8	12	3.1%

Table 3-2: Comparisons of number of requests in expMi and expEn sessions.

ExpMi sessions were short. In terms of time, 20% were less than two seconds and 51.6% were less than 30 seconds. In terms of requests, 73.9% of sessions were the bare minimum of two requests and 92.8% of sessions consisted of just two or three requests.

ExpMi sessions were compared with expEn sessions to see if the sessions were similar in length. Initial glances at both tables seem to indicate that the expMi and expEn sessions were similar in length, both in duration and number of requests. A chi square statistical analysis undertaken on the length of sessions in seconds gives a p value of $1.9 * 10^{-5}$ and in requests gives a p value of $3.5 * 10^{-3}$. These values indicate that there was a significant statistical difference in the duration (in time) between expMi and expEn sessions but not a statistical difference in the number of requests. These values also suggest that clients in expMi sessions spent a significantly longer time reading the interface pages of the Niupepa web site than clients in expEn sessions. This supposition can be confirmed by examining how long it took expMi clients and expEn clients to read interface pages of the Niupepa digital library.

3.3.2 How Long to Read Interface Pages?

The length of time from when a client requests a page of the interface to when the client requests another page of the interface, the apparent page read time, was calculated. Since it was not possible to confirm that the clients were only reading the pages of the interface during this time, these times have been labelled apparent page read times. These times give indications about how long clients take to read interface pages.

The apparent page read times were grouped in intervals of 1–2 seconds, 3–9 seconds, 10–19 seconds, 20–39 seconds, 40–79 seconds and 80+ seconds. The apparent page read times for both expMi and expEn sessions are displayed in *Table 3-3*.

Apparent Read Times in Exploratory Sessions

Seconds	expMi	expEn
Total Apparent Reads	246	491
1–2	55	59
3–9	36	101
10–19	33	89
20–39	37	88
40–79	21	57
80+	64	97

Table 3-3: expMi and expEn apparent read times of Niupepa interface pages in exploratory sessions.

A chi square analysis was undertaken on the apparent read times and returned a p value of $5.6 * 10^{-4}$. This value indicates that there was a significant statistical difference in the expMi and expEn apparent read times of interface pages. To determine exactly what this difference was, statistical summaries of the apparent read times were calculated and are displayed in *Table 3-4*.

Statistical Summary for Exploratory Apparent Read Times

	expMi	expEn
Total Apparent Reads	246	491
Average (seconds)	145.6	106.9
Standard Deviation (seconds)	410.1	339.7
Mode (seconds)	0	0
Shortest session (seconds)	0	0
Lower Quartile (seconds)	4	8
Median (seconds)	19	19
Upper Quartile (seconds)	74	53
Longest session (seconds)	3396	3497

Table 3-4: Statistical summaries comparing apparent read times in expMi and expEn exploratory sessions.

These statistics displayed in *Table 3-4* show that the average apparent read time of expMi sessions was almost 2.4 minutes (145.6 seconds). The average apparent read time of expEn sessions was 1.8 minutes (106.9 seconds). Both medians are equal at 19 seconds but the higher upper quartile of expMi suggests that these sessions had a higher proportion of longer exploratory sessions.

These statistics suggest that exploratory clients of the Niupepa digital library who used the interface in te reo Māori took more time to read the interface pages than clients who used the interface in the English language.

3.3.3 Which Interface Pages Were Requested?

The interface pages requested by expMi and expEn clients are listed in *Table 3-5*. Note that an interface page refers to a page of the Niupepa web site; it does not refer to a newspaper document in the Niupepa collection.

	expMi		expEn	
Total pages	426	100.0%	878	100.0%
Niupepa home page	358	84.0%	728	82.9%
NZDL home page	9	2.1%	60	6.8%
About English Abstracts	38	8.9%	35	4.0%
help page	5	1.2%	24	2.7%
preferences page	15	3.5%	29	3.3%
about greenstone page	0	0.0%	1	0.1%
unknown page	1	0.2%	1	0.1%

Table 3-5: *The numbers and percentages of web site interface pages requested in exploratory sessions by expMi and expEn.*

The majority (84.0%) of expMi session requests were requests for the Niupepa home page, including links to different sections on the home page. This preference is to be expected because the home page is usually the first page accessed by clients exploring a web site. The second highest preference for the

expMi sessions, at 8.9%, was the About English Abstracts page, which initially suggests that expMi users are interested in reading about the English Abstracts. However, the term used in the link to this page, ‘he rāpopototanga Reo Pākehā’, could also be translated as ‘An English Summary’. Hence, expMi users could have been mistakenly accessing this link in the hope of viewing a summary of the Niupepa collection in English. By contrast, the second highest preference in expEn sessions, at 6.8%, is the NZDL home page, suggesting that exploratory clients in English were also interested in information about the NZDL.

3.3.4 Summary—Exploratory Sessions

Exploratory sessions in te reo Māori (expMi) were short sessions, mostly consisting of two or three requests, and half of the sessions were completed within 30 seconds. While the median apparent read times of interface pages in expMi sessions was 19 seconds, the average apparent read time was 146 seconds, suggesting that there were many expMi sessions where the clients were taking longer than two minutes to read and explore the Niupepa interface in te reo Māori. The home page, and sections within the home page, was the interface page mostly requested by expMi users. The About English Abstracts explanation page was also significantly requested by expMi sessions; however, this may have been a result of an ambiguous translation of this hyperlink.

Exploratory sessions only request pages from the Niupepa interface; no documents are requested and no queries are submitted. Aside from the details listed in this summary, it is difficult to derive any further information about how te reo Māori was used by exploratory (only) clients of the Niupepa digital library.

This summary leads us into the next part of the analysis—how te reo Māori was used by clients who were actively using the Niupepa digital library to access information. Usage of the Niupepa digital library by clients in te reo Māori (usgMi clients) will be analysed in three parts, interface behaviours, document access behaviours and query behaviours. The first part, an analysis of how usgMi clients used the Niupepa interface, begins in the following section.

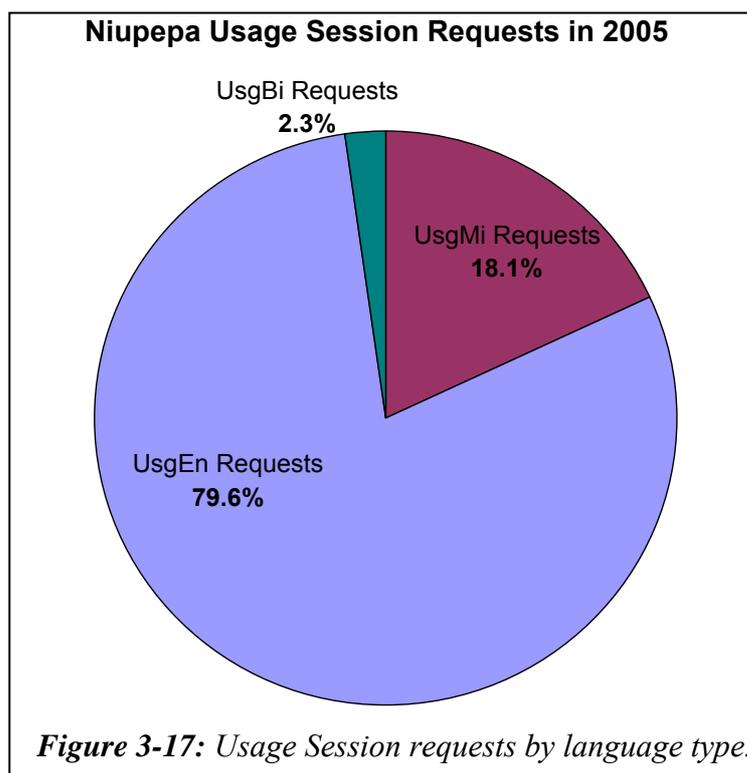
3.4 Niupepa Interface Behaviours

This section analyses how the Niupepa interface was used by reo Māori clients in usage sessions. A usage session is a session where the client actively engages with the Niupepa web site to view documents in the Niupepa collection (for a description of how usage sessions are defined see §2.4.2). Usage session data is the most important data to analyse as it explicitly examines the key target group, clients who actively use the Niupepa digital library.

There were 172,905 usage requests recorded in 6,714 sessions for the Niupepa collection in 2005. Of these requests, 33,540 (19.4%) were when the interface was set to reo Māori and 139,365 (80.6%) were when the interface was set to English. Usage sessions lasted approximately ten times longer and had approximately ten times more requests than the exploratory sessions.

The usage session requests were then subdivided by language type (as explained in §2.4.3). There were 31,287 requests in 1,702 usage sessions when the interface language was mostly set to reo Māori (usgMi). There were 137,697 requests in 4,952 usage sessions when the interface language was mostly set to English (usgEn). There were 3,921 requests in 90 usage sessions when the interface language was defined as bilingual (usgBi). A count of the usgMi requests shows the interface language set to te reo Māori in 98.5% of the requests. A count of the usgEn requests shows the interface language set to English in 98.8% of the requests. A count of the usgBi requests show the interface set to English in 73.5% of the requests and set to reo Māori in the remaining 26.5%.

The usage session requests for each language type were calculated as percentages and are displayed in *Figure 3-17*.



Statistics generated from TLA were used to examine seven interface behaviours exhibited by clients who used the Niupepa digital library in te reo Māori. The interface behaviours examined were:

- the length of usgMi sessions
- the number of return usgMi sessions
- the origin of usgMi sessions
- the interface pages requested in usgMi sessions
- the apparent page read times of interface pages in usgMi sessions
- the frequency of interface language switching in usgMi sessions
- the position of interface language switching in usgMi sessions.

These seven interface behaviours comprise the next seven sub-sections and provide a description of how te reo Māori clients used the Niupepa interface.

3.4.1 How Long Were Usage Sessions?

Two measurements were used to establish the length of usgMi sessions; duration in time (seconds and minutes) and length in number of requests. Statistics

generated from TLA on the Niupepa digital library are listed to show how long usgMi sessions of the Niupepa digital library were in 2005. These statistics are listed alongside length statistics of usgEn sessions and usgBi sessions.

How Long in Time Were Usage Sessions?

UsgMi sessions, like all usage sessions, generally did not last for a long time. The shortest usage sessions lasted less than a second, the longest for almost six hours and the average was 20 minutes. The median time for usage sessions was 6 minutes.

The durations of usgMi sessions, usgEn sessions and usgBi sessions were grouped into the following time intervals; 0–10 seconds, 11–60 seconds, 1–3 minutes, 3–15 minutes, 15–30 minutes, and more than 30 minutes. The numbers of sessions that fitted into these time categories were calculated as percentages for comparison purposes and graphed in *Figure 3-18*.

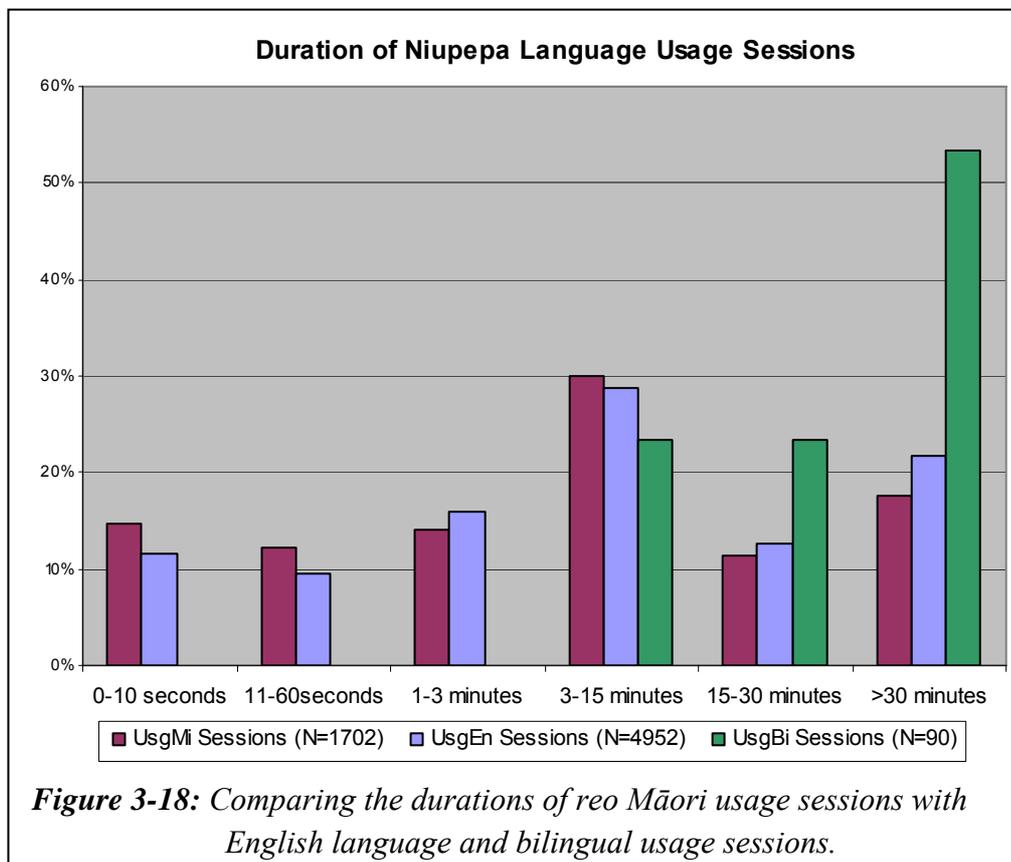


Figure 3-18 shows that usgMi sessions and usgEn sessions appeared to have similar session durations, while the usgBi sessions were all longer sessions. However, when chi square analysis is undertaken on the session lengths of usgMi and usgEn sessions it generates a p value of $3.0 * 10^{-6}$. This indicates that the difference is statistically significant. UsgMi sessions had a higher proportion of shorter sessions (in the 0–10 second and 11–60 second category) while usgEn sessions had a higher proportion of longer sessions (in the 15–30 minute category and >30 minute category). The average usgMi session duration was 16.8 minutes; the average usgEn session duration was 21.3 minutes. Although *Figure 3-18* appears to show that usgMi sessions were of similar duration to usgEn sessions, the figures indicate that usgMi sessions were statistically shorter than usgEn sessions.

The statistics for usgBi sessions indicated that these sessions were almost three times longer in average length than usgMi sessions. One of the criteria for determining a usgBi session was that these sessions must be at least 10 requests long. This criterion alone ensures that usgBi sessions are long. It is also difficult to deduce usgBi characteristics with the same amount of confidence as usgMi and usgEn characteristics because the number of sessions is comparatively low at only 90 sessions.

How Long in Number of Requests Were Usage Sessions?

Usage sessions consisted of requests for pages from the Niupepa interface and at least one request for an actual document from the Niupepa collection. The shortest usage sessions involved just two requests and sessions of just two requests were the most common length. The longest session involved 598 requests. There was an average of 25.6 requests per usage session.

To illustrate the requests recorded in usgMi sessions the numbers of requests per usage session were grouped into the following divisions; 2 requests, 3–4 requests, 5–8 requests, 9–20 requests, 21–40 requests and 41+ requests. The numbers of sessions per request count category were calculated as percentages and are displayed in *Figure 3-19*.

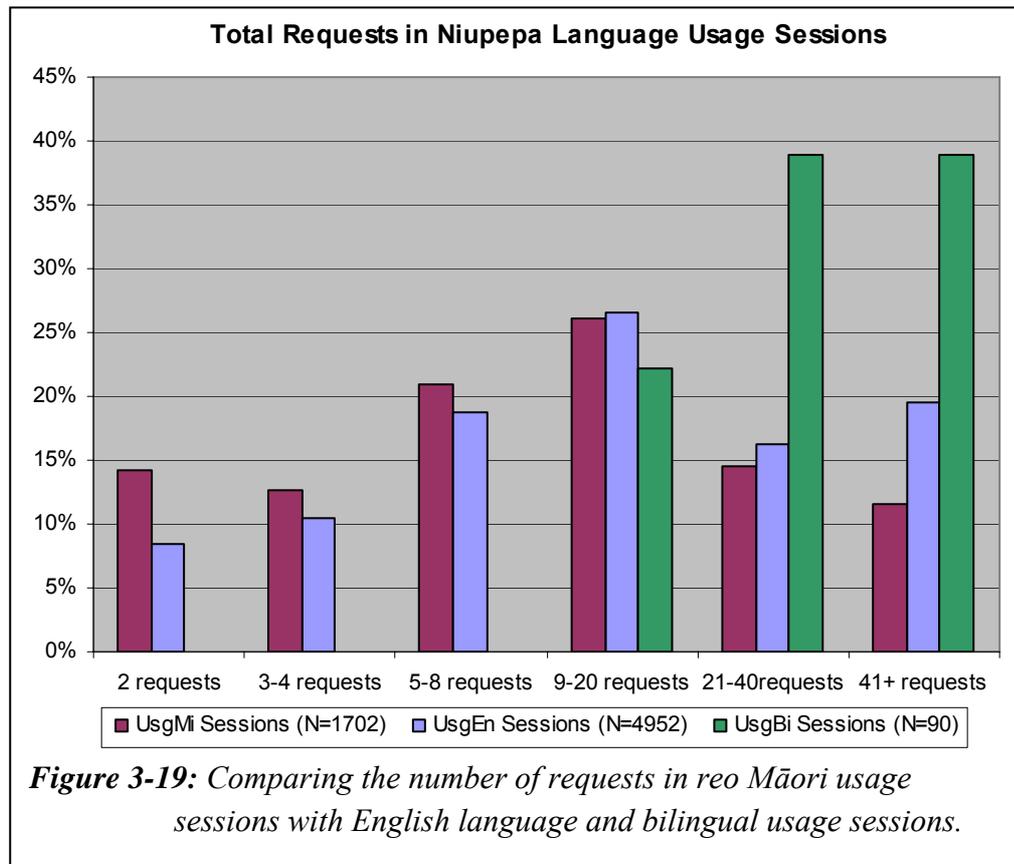


Figure 3-19 shows that usgMi sessions had the highest percentages of short sessions; sessions with only 2 requests, 3–4 requests and 5–8 requests. The usgEn sessions had a higher proportion of longer sessions than the usgMi sessions but the usgBi sessions had the highest proportion of longer sessions.

UsGmi sessions had an average of 18.4 requests per session and a median of 9 requests per session. The usgEn sessions had an average of 27.8 requests per session and a median of 12 requests per session. Thus the usgMi sessions, on average, consisted of a third fewer requests than the usgEn sessions.

The statistics for usgBi sessions indicated that these sessions had almost twice as many requests as usgMi and usgEn sessions. Again, this high number of requests may simply be because of the criteria for determining a usgBi session, that these sessions must be at least 10 requests long.

3.4.2 How Long to Read Interface Pages?

Reading of interface pages has been defined as apparent page reading (see §2.5.1). The apparent page read time is the length of time from when one interface page of the web site is requested until another interface page or document is requested. The apparent read times in usgMi, usgEn, usgBi sessions and in all usage sessions were calculated and descriptive statistics have been displayed in *Table 3-6*.

UsgMi clients, on average, took 51.5 seconds to read the interface pages in usage sessions. This apparent page read time is almost 1/3 faster than it took expMi clients to read interface pages (145.6 seconds) in exploratory sessions (*Table 3-4*).

	usgMi	usgEn	usgBi	Total
Count	1339	5810	542	7691
Average (seconds)	51.5	48.7	71.8	50.8
Standard Deviation	162.1	187.2	223.0	186.0
Mode (seconds)	4	4	4	4
Shortest (seconds)	0	0	0	0
Lower Quartile (seconds)	5	6	6	6
Median (seconds)	10	12	15	12
Upper Quartile (seconds)	31	32	40	33
Longest (seconds)	2131	3237	2367	3237

Table 3-6: Summary statistics comparison of usgMi, usgEn, usgBi and total apparent page read times of interface pages in usage sessions.

The average apparent usgEn page read time (48.7 seconds) seems similar to the average usgMi apparent read time (51.5 seconds). However a chi square analysis comparing the usgMi and usgEn apparent read times gives a p value of $4.8 * 10^{-11}$, indicating that the usgMi apparent page read times are significantly longer.

3.4.3 How Many Clients Returned?

TLA of the Niupepa web log can give indications of how often clients returned to the Niupepa digital library. The results generated assume that clients always have cookies enabled, the cookies are not routinely deleted, clients always log on using the same account on the same computer, and that there is minimal to nil network and local browser caching involved. These assumptions are not valid for some users in some sessions. However, analysing return statistics can still give an indication of how often usgMi clients returned to use the Niupepa digital library.

Statistics from TLA indicated that from a total of 6,117 identifiable clients there were at least 283 (4.6%) clients who undertook return sessions. These 283 clients made 33,099 (19.1%) requests for pages and documents from the Niupepa web site. These figures indicate that only a small percentage of clients were identified as return clients but these clients generated a high proportion of the requests.

Of the usgMi clients 82 (5.2%) undertook return sessions requesting 4,316 (13.8%) pages and documents. There were also 192 (4.3%) usgEn clients who undertook return sessions requesting 27,112 (19.7%) pages and documents. There were 52.6 requests submitted per returning usgMi client compared with 141.2 requests submitted per returning usgEn client, indicating that the return usgEn clients submitted more requests than the return usgMi users.

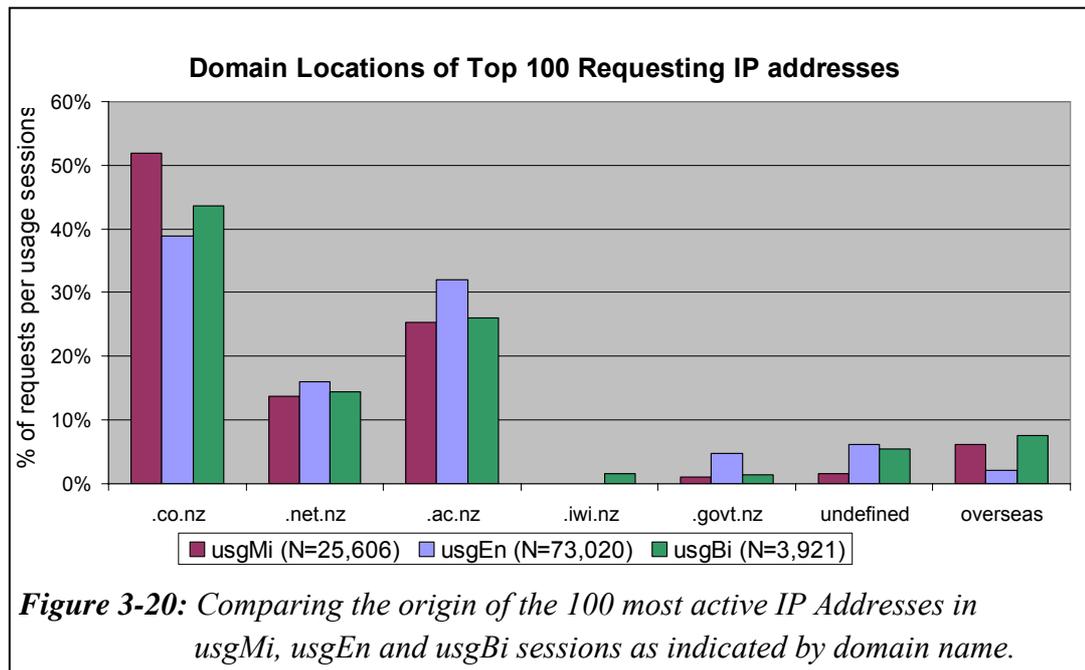
None of the clients classified as usgBi clients undertook return sessions as a usgBi client but 13 (14.4%) of these clients were recorded as other user types. There were two usgBi clients who were also recorded as usgMi clients and 11 usgBi clients who were also recorded as usgEn clients. These figures suggest two usage behaviours. First, switching the language of the interface is an activity undertaken only once by users. Second, a proportion of usgEn clients are prepared to try the interface in the medium of te reo Māori.

3.4.4 Where Did Usage Sessions Originate?

Information about the origin of requests can be derived from TLA. The second argument recorded in the web logs with each request was the unique IP Address (Internet Protocol address) of the computer or network device that submitted the request. The IP Addresses can be examined by a process called reverse DNS lookup, which returns information about the computer/network device including domain name and location. Counts can then be undertaken to determine the number of requests submitted per domain name and per location.

The IP Addresses of the requests recorded in usgMi sessions were examined using reverse DNS lookup to see if any trends in domain name could be determined. Eight distinctive domain name types were identified, `.co.nz`, `.net.nz`, `.ac.nz`, `.iwi.nz`, `.govt.nz`, *undefined* and *overseas IP*. The first six have a common first level domain name (i.e. `.nz`) but are different in the 2nd level domain name; the seventh domain type, *undefined*, represents those IP Addresses which did not return a valid domain name when queried; and the eighth domain type is a grouping of all the first level domain names which did not equal `.nz` and were subsequently grouped as being requests from overseas.

In total 3,183 IP Addresses submitted requests to the Niupepa digital library in 2005. Determining individual domain names on all 3,183 IP Addresses is beyond the scope of this thesis. Consequently, only the 100 IP Addresses submitting the most requests in usgMi sessions were examined. To provide a comparison the 100 IP Addresses submitting the most requests in usgEn sessions and the total 69 IP Addresses making requests in the usgBi sessions were also examined. The percentages of requests in these usage sessions are displayed in *Figure 3-20*.



The usgMi requests had a similar origin to the usgEn and usgBi requests; most requests came from the .co.nz domain, followed by the academic community the .ac.nz domain, with the third largest domain being the .net.nz group. These results suggest that the use of te reo Māori in the Niupepa digital library was not limited to one particular domain type.

The .co.nz group has been defined as organisations pursuing commercial aims and purposes (Office of the Domain Name Commissioner, 2005) and the .net.nz group has been defined as organisations and service providers directly related to the New Zealand (Aotearoa) Internet. The two domain names have been used for similar purposes by Internet Service Providers in Aotearoa and if these two names are grouped together they account for an average of 59.6% of the IP Addresses making requests in each usage session group.

The .ac.nz domain name is the origin of the second largest group of IP Addresses and is analysed in more depth in the following sub-section. The IP Addresses which indicate requests from overseas also reveal some interesting information, as discussed in a following sub-section.

There was one IP Address that originated from the second level domain name of `.iwi.nz`. This domain name is reserved for traditional Māori tribes. It seems unusual that this domain name only appears in the usgBi sessions. Perhaps if more than the top 100 IP Addresses were examined in the usgMi and usgEn sessions then the `.iwi.nz` second level domain name may have appeared in those sessions as well.

If we examine the second level domain name of `.govt.nz` we see that there are three times as many IP Addresses making request in English usage sessions as there are in te reo Māori or bilingual usage sessions. This is, perhaps, an indication of a low number of reo Māori speakers working in the New Zealand Government.

Originating Academic Institutions

The requests recorded from the academic second level domain name were isolated and grouped by the originating academic institution. These academic institutions are displayed in *Figure 3-21* sorted by location and listed from north to south.

Nearly half of all Māori live in Auckland, Waikato and the Bay of Plenty (Te Puni Kōkiri, 2001b). As the Niupepa collection is primarily a reo Māori resource we would expect most of the requests to come from this area and this expectation is indeed reflected in *Figure 3-21*.

Statistics provided by the Ministry of Education for 2005 (Alosius, 2007) show that 3,613 Māori students enrolled at Massey University, 3,028 Māori students at Waikato Institute of Technology, 2,802 Māori students at the University of Auckland and 2,239 Māori students at the University of Waikato. These figures represent the four institutions displayed in *Figure 3-21* that have the highest number of Māori students enrolled. The Eastern Institute of Technology had the highest percentage of Māori student enrolments at 22.5%, followed by Waikato University at 16.7%, the Southern Institute of Technology at 14.9% and the

Waikato Institute of Technology at 14.7%. All other education providers had less than 10% of Māori student enrolments.

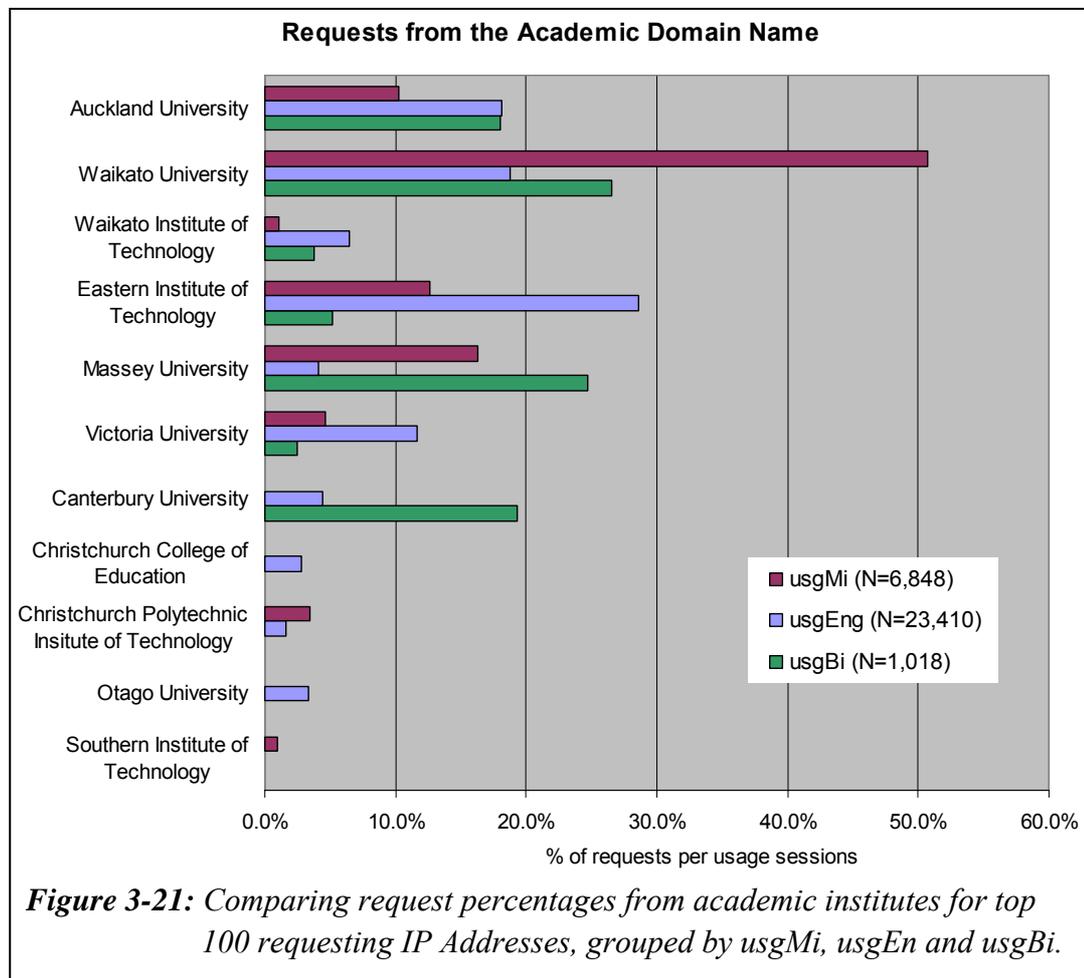


Figure 3-21 indicates more requests coming from the northern tertiary institutions, Auckland University down to Massey University. This trend mirrors the statistics provided by the Ministry of Education which state that two thirds (65.5%) of the Māori student population is studying at these institutions.

There was a proportionally higher number of requests originating from the University of Waikato. As Waikato University is the host institute of the Niupepa digital library perhaps the digital library is more widely known and, as a result, more widely used at this university.

Originating Countries

The requests from 1st level domain names that did not equal .nz were examined to determine which countries were making the most requests to the Niupepa collection. Overseas IP Addresses that made 20 or more requests to the Niupepa collection were collated and then plotted in *Figure 3-22*. It should be noted that requests from Hawai‘i were separated from the US requests even though Hawai‘i is a state of the United States.

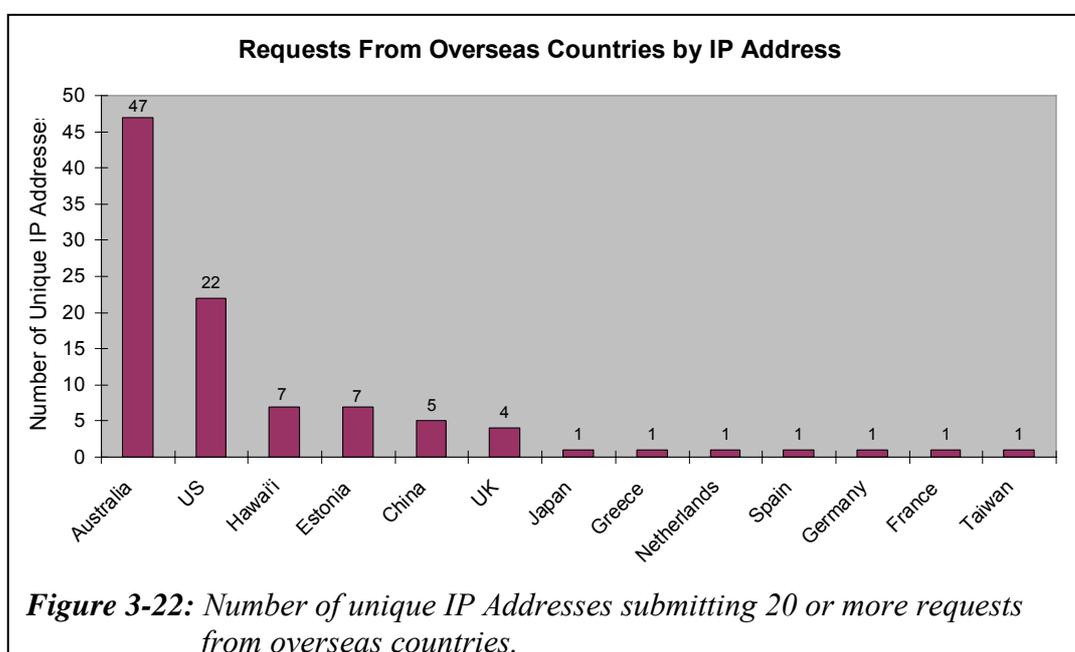


Figure 3-22 shows some interesting characteristics. First, it highlights that the most overseas usage requests in 2005 were from Australia, with 47 separate IP Addresses (submitting 2,697 usage requests). Given that estimates as early as 2001 suggested over 90,000 Māori (over 15% of all Māori) live in Australia (Bedford et al, 2004) this high interest in the Niupepa web site by people in Australia, presumably Māori living there, is not surprising. Following Australia, the next country using the collection was the US, with 22 unique IP Addresses submitting 847 usage requests. Next was Hawai‘i, a state of the US, with 7 separate IP Addresses submitting 562 usage requests. A significant number of requests from Hawai‘i is not surprising given that the Hawaiian and Māori people

are both Pacific Island nations sharing a similar language and culture, and that both nations have been involved with building digital libraries based on legacy newspapers.

Perhaps what is surprising though is the reasonably high appearance in the graph of Estonia, with 7 unique IP Addresses making 441 requests in usage sessions. As 403 of those requests occurred in April and May, one can speculate that these Estonian usage sessions represent a particular concentrated effort from a person, or group of people working together, to research some aspect of Aotearoa's encounter history, some aspect of Aotearoa's indigenous language or history or, perhaps, some aspect of using the Greenstone software to build digital libraries.

Most (82.0%) of the usage requests displayed by IP Addresses in *Figure 3-22* are requests from usgEn sessions. There were also 14.8% recorded from usgMi sessions and the remaining 3.2% were from usgBi sessions. Countries with 75% or more of requests in English language sessions included Australia, US, Estonia, China, UK, Japan, Spain, Germany and Taiwan. Countries that appeared prominent in the reo Māori usage sessions were Hawai'i, with 351 requests representing 62.5% of its usage requests and Greece, with 31 requests at 64.6% of its usage requests. In addition, the 21 requests from the solitary French IP Address were all recorded as reo Māori usage session requests.

Only one country, the Netherlands, appeared prominent in bilingual usage session requests with 30 (65.2%) of its 46 requests recorded as bilingual usage sessions and the other 16 being recorded as English usage sessions. These figures suggest that clients from Hawai'i, the client from Greece, and the client from France were willing to undertake activities on the Niupepa digital library in te reo Māori. The figures also suggest that the client from the Netherlands switched languages often when using the web site.

3.4.5 Which Interface Pages Were Requested?

The interface pages requested by clients who use te reo Māori are examined in this sub-section. Note: a page refers to an interface page of the Niupepa web site rather than a newspaper document within the Niupepa collection.

The numbers and percentages of interface pages requested in usgMi sessions are listed in *Table 3-7*. Also listed for comparative purposes are the numbers and percentages of pages requested in usgEn sessions and usgBi sessions.

	usgMi		usgEn		usgBi	
Total pages	1390		6080		551	
Niupepa home page	1127	81.1%	4931	81.1%	413	75.0%
NZDL home page	18	1.3%	121	2.0%	9	1.6%
About English Abstracts	83	6.0%	349	5.7%	26	4.7%
help page	48	3.5%	150	2.5%	25	4.5%
preferences page	114	8.2%	529	8.7%	78	14.2%
about greenstone page	0	0.0%	0	0.0%	0	0.0%
unknown page	0	0.0%	0	0.0%	0	0.0%

Table 3-7: The numbers and percentages of interface pages requested by usgMi, usgEn and usgBi clients.

Table 3-7 shows the majority of usgMi requests, at 81.1%, were requests for the Niupepa home page which, as a home page, can be expected. While the second most requested interface page on the expMi (exploratory) sessions was the About English Abstracts page, here in usage sessions the second most requested page was the Preferences page at 114 (8.2%) requests.

This characteristic is also expected when clients are actively using the collection as they are likely to request the Preferences page to alter language, format and searching preferences. However, while this is a significant proportion of the interface pages requested in usgMi sessions, it is not a significant proportion of

the total requests. There were 31,287 total requests in usgMi sessions; 114 requests for the Preferences page represent less than 0.4% of these total requests.

The percentages of interface pages accessed by reo Māori and English sessions were similar. When chi square analysis is undertaken on the counts of usgMi and usgEn interface pages a p value of 0.11 is generated. This value indicates that there is no significant difference in the type of interface pages that are being requested in usgMi and usgEn sessions. Thus we can conclude that clients who used the Niupepa digital library in te reo Māori were requesting the same proportion of interface pages as those who used the Niupepa digital library in English.

3.4.6 How Often Was the Interface Language Switched?

In general, usgMi clients showed a reluctance to switch the language of the interface. Of 1702 usgMi sessions only 142 (8.3%) switched the interface language; most of these switches, 114 of 193 (59.1%), were a switch from English to reo Māori.

By comparison, 1067 (21.6%) of 4952 usgEn clients switched the interface language; most of these switches, 1058 of 1244 (85.1%), were a switch from reo Māori to English. By definition a usgBi session contains 3 or more language switches; there were 90 usgBi sessions involving 352 interface language switches, 154 (43.7%) of these were a switch from English to reo Māori and 198 (56.3%) a switch from reo Māori to English.

3.4.7 When Was the Interface Language Switched?

The usgMi sessions were divided into 10 equal portions and the portion in which the language switch occurred in each session was noted. These positions are plotted in *Figures 3-23* and *3-24*. The positions that the language switch occurred in usgEn and usgBi sessions are also plotted in *Figures 3-23* and *3-24*.

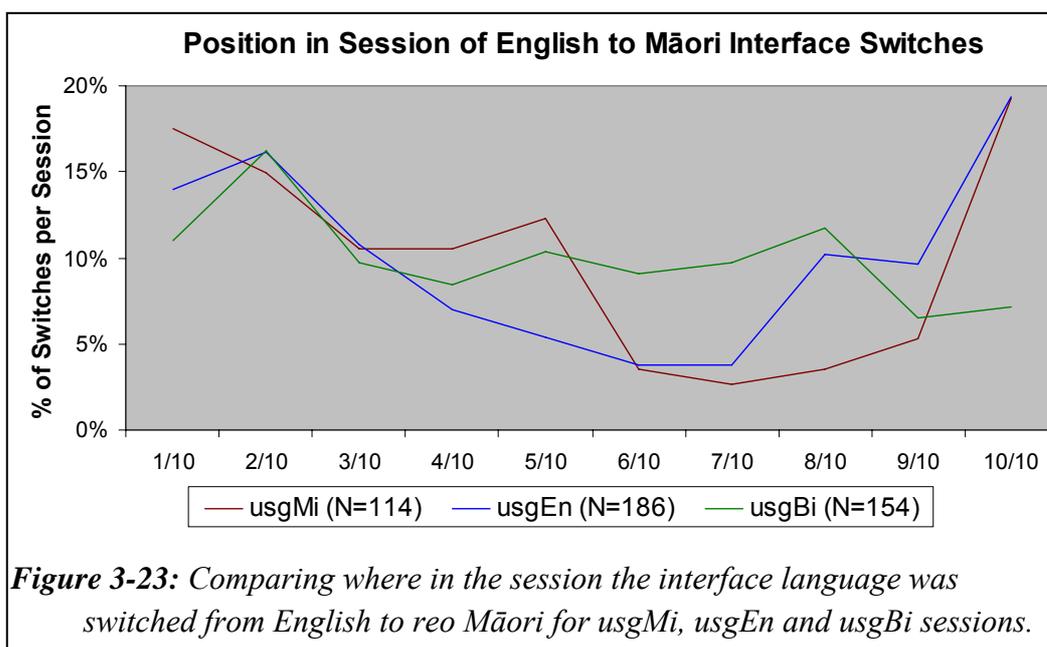
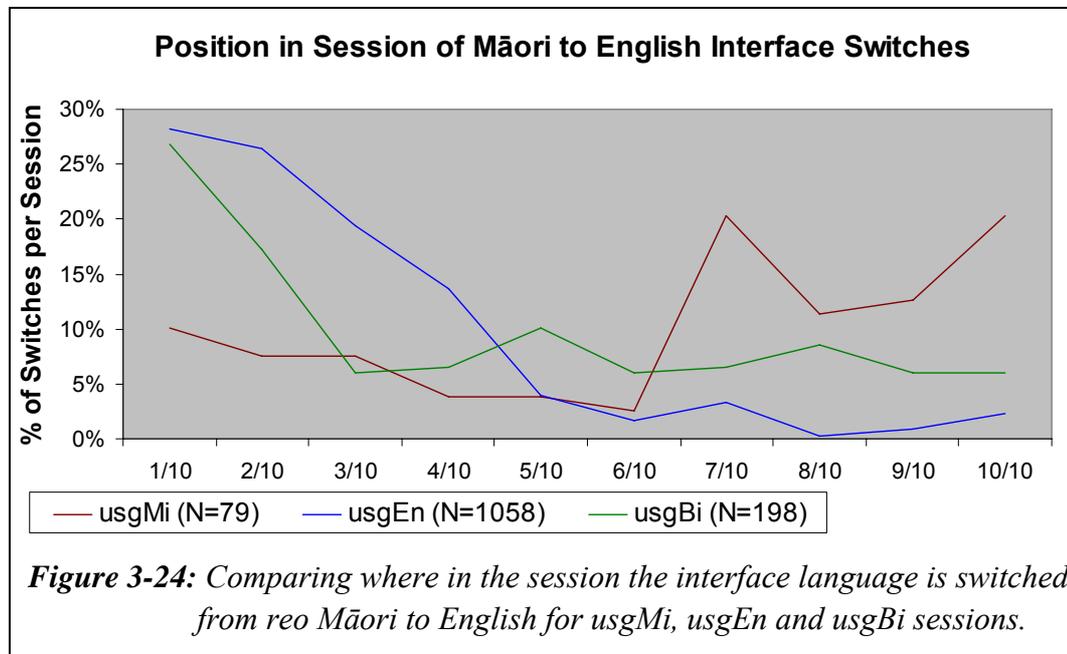


Figure 3-23 shows that clients in usgMi sessions mostly switched the interface language from English to reo Māori in the first half of the session or the very last tenth of the session. There were only 114 interface language switches to reo Māori in usgMi sessions. This low number of switches to reo Māori by usgMi clients suggests that many of the 1702 usgMi sessions were occurring because either the default language of the web site was set to reo Māori or the clients were accessing the web interface through a bookmark that included the default interface language setting of reo Māori.

The usgEn sessions had a higher number (186) of interface switches into reo Māori than the usgMi sessions (114). This number indicates that many usgEn clients must be confident enough in their reo Māori abilities to try the interface in te reo Māori. As most of these switches occurred in the first 3/10ths of the

sessions it perhaps indicates usgEn clients exploring the reo Māori version of the interface.



In contrast, *Figure 3-24* shows the switch of the interface language from reo Māori to English by usgMi clients occurs mostly in the final 4/10ths of usgMi sessions. The position of this switch could suggest a frustration in the ability to use the interface in te reo Māori, although the number of switches is low, 79, representing less than 5% of the total 1702 usgMi sessions.

The largest number of interface switches was from reo Māori to English in usgEn sessions. There were 1058 switches to English in the 4952 usgEn sessions. These figures indicate that more than 1 in 5 usgEn clients switched the interface language to English. As these switches were mostly undertaken at the start of a session it suggests that setting the default language of the Niupepa interface to English would be welcomed by clients who use the Niupepa digital library in English. The default language setting of the web site is comprehensively discussed in Chapter 5.

3.4.8 Summary—Interface Behaviours

This section analysed how the Niupepa interface was used by te reo Māori clients in usgMi sessions. Statistics drawn from TLA examined seven aspects of using the Niupepa interface in te reo Māori.

The statistics reveal that on average a usgMi session lasted for 30 minutes and consisted of 18 requests. The average apparent read time of an interface page by usgMi clients was 52 seconds. Only 6% of the usgMi sessions were identified as return sessions; however, these return sessions were responsible for 14% of the total usgMi requests. More than half of the usgMi requests from an academic institution in Aotearoa originated from Waikato University. There were twice as many usage requests from Australia as there were from any other overseas country. The home page was the most popular interface page requested by te reo Māori clients, at 81% of all interface page requests, followed by the Preferences page at 8%. The language of the interface was only switched 142 times in 1702 usgMi sessions. When the language of the interface was switched, it was switched more often from English to reo Māori (59%) and mostly in the first half of the session.

This brief summary completes the examination of how the Niupepa interface was used by clients in te reo Māori. Usage analysis of the Niupepa digital library by usgMi clients was undertaken in three parts; an analysis of how the interface was used, of how documents were accessed and of how queries were undertaken. Having completed this analysis the next stage is to examine how documents were accessed by usgMi clients. This analysis begins in the next section.

3.5 Document Access Behaviours

This section analyses how te reo Māori clients accessed Niupepa documents in usage sessions. By definition (see §2.4.2) if a client requested a document in a session then that session was regarded as a usage session. A document request is defined as a request for some document within the Niupepa digital library that represents real content, rather than something that provides access to content, such as an interface page. Typically this resource is a newspaper page but it could also be a request for an English Abstract (for an example see *Figure 3-9*) or a Bibliographic Commentary (for an example see *Figure 3-10*).

TLA was used to provide statistics on seven document access behaviours exhibited by clients who used the Niupepa digital library in te reo Māori. The document access behaviours examined were:

- the methods used to access documents in usgMi sessions
- the document formats favoured in usgMi sessions
- the document language favoured in usgMi sessions
- the publication years most requested in usgMi sessions
- the propensity for documents with associated text files
- the propensity for documents with an associated English Abstract
- the apparent read times of documents by usgMi clients.

These seven document access behaviours comprise the next seven sub-sections and provide a thorough description of how te reo Māori clients accessed documents of the Niupepa digital library.

3.5.1 How Were Documents Accessed?

There were three methods available to access documents of the Niupepa digital library; selection of a document following a query activity, selection of a document from a list of publication series, and selection of a document from a list

based on dates. When a client requested a document by one of these methods, say a query, and then proceeded to browse documents in the collection, the subsequently browsed documents were all recorded as being selected from the query activity.

Results from TLA indicated the methods used by usgMi clients to access Niupepa documents. The counts and percentages are displayed in *Table 3-8*, along with counts and percentages of how documents were accessed by usgEn and usgBi clients.

How Documents Were Accessed								
	usgMi		usgEn		usgBi		Total	
Total	22,248	100%	102,040	100%	2,272	100%	151,260	100%
From query	17,282	77.7%	83,221	81.6%	1,595	70.2%	102,098	80.7%
From series	2,878	12.9%	6,754	6.6%	233	10.3%	9,865	7.8%
From date	1,641	7.4%	9,860	9.7%	346	15.2%	11,847	9.4%
Uncertain	447	2.0%	2,205	2.2%	98	4.3%	2,750	2.2%

Table 3-8: Counts and percentages comparing how documents were accessed by usgMi, usgEn and usgBi clients and the total usage session clients.

Table 3-8 shows that the query activity (see *Figure 3-3*) was the most popular method used by usgMi clients to access Niupepa documents. The usgMi clients showed a second highest preference for accessing documents from the series list (see *Figure 3-4*) and the least popular method was accessing the document from a list based on dates (see *Figure 3-5*).

Clients in all three usage sessions preferred to access documents from a query activity. However, the usgEn and usgBi clients differed from the usgMi clients in that the second preferred method for accessing documents was from a list based on dates, while the least preferred method was from the series list.

3.5.2 Which Document Formats Were Favoured?

The newspaper pages stored in the Niupepa digital library can be presented to the client in one of three formats; extracted text generated from the OCR process, a preview image and a full size image. Examples of these formats are shown in *Figures 3-6, 3-7, and 3-8*. *Table 3-9* lists the numbers and percentages of these three formats, along with a fourth undefined listing where the format of the page was not recorded correctly in the web log. The undefined requests are most likely caused by bookmarked pages being requested following software changes to the Niupepa web site.

Formats of Requested Documents						
	usgMi		usgEn		usgBi	
Total Documents	22248	100%	102040	100%	2272	100%
Extracted Text	13730	61.7%	56042	54.9%	1238	54.5%
Full Size Image	7389	33.2%	39052	38.3%	816	35.9%
Preview Image	999	4.5%	6379	6.3%	211	9.3%
Undefined	130	0.6%	567	0.6%	7	0.3%
Chi square p value	4.3 * 10 ⁻⁸⁰					
Chi square p value	6.8 * 10 ⁻⁸⁷					

Table 3-9: Document formats requested by usgMi, usgEn and usgBi clients.

The document format preferred by usgMi clients was the extracted text format, at 61.7% of all requested newspaper documents. The next preferred format was the full size image at 33.2% and the preview image (4.5%) was the least requested format.

Table 3-9 also shows that all three usage types prefer the Extracted Text format of the document. While the percentages do appear reasonably similar, chi square analysis shows that there is a significant statistical difference in usgMi and usgEn document formats (p value = 4.3 * 10⁻⁸⁰) and also a significant statistical difference in usgMi, usgEn and usgBi document formats (p value = 6.8 * 10⁻⁸⁷).

These p values indicate the usgMi sessions have a higher preference that is statistically significant to request the extracted text formats, while the usgEn and usgBi have a higher statistically significant preference to request the preview and full size image formats.

3.5.3 Which Document Types Were Favoured?

The documents that were available to Niupepa clients consisted of:

- 19,106 newspaper pages written in te reo Māori
- 15,696 newspaper pages written in te reo Māori and English
- 680 newspaper pages written in English
- 313 English Abstract documents that summarise 10 of the newspaper periodicals
- 22 Bibliographic Commentaries written in English with te reo Māori quotations.

To determine which types of documents were most popular to new and general clients, and to remove the influence of heavy users, the top 100 documents requested by unique users were counted. These documents were requested a total of 5,681 times; 1,186 by usgMi clients, 4,219 by usgEn clients and 276 by usgBi clients. The usage session percentages for each type of document count are displayed in *Figure 3-25*.

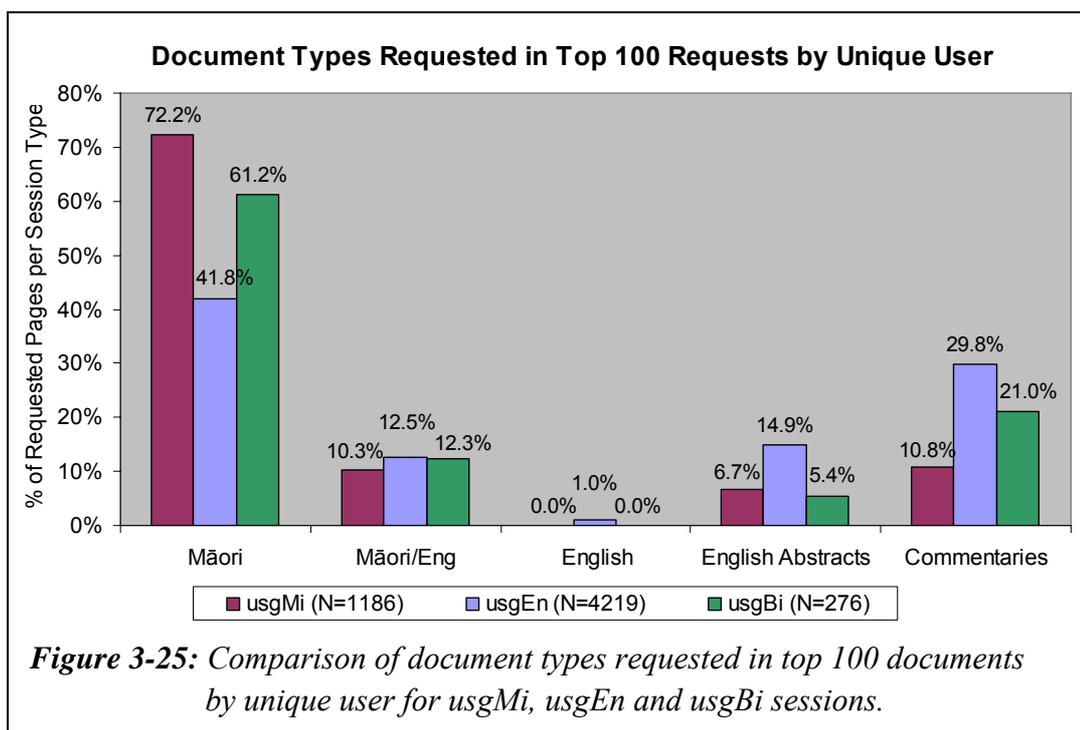


Figure 3-25 shows usgMi clients have the highest percentage of requesting newspaper documents in te reo Māori; 72.2% of the top 100 documents requested by usgMi clients were Niupepa pages written in te reo Māori, compared with 61.2% from usgBi clients and a low 41.8% from usgEn clients.

The usgEn clients had the stronger preference to access documents written in English; the commentary information (29.8%), the English Abstracts (14.9%) and the newspaper pages written in English (1.0%). UsgBi clients requested documents that were between the extremes of those requested by usgMi clients and usgEn clients.

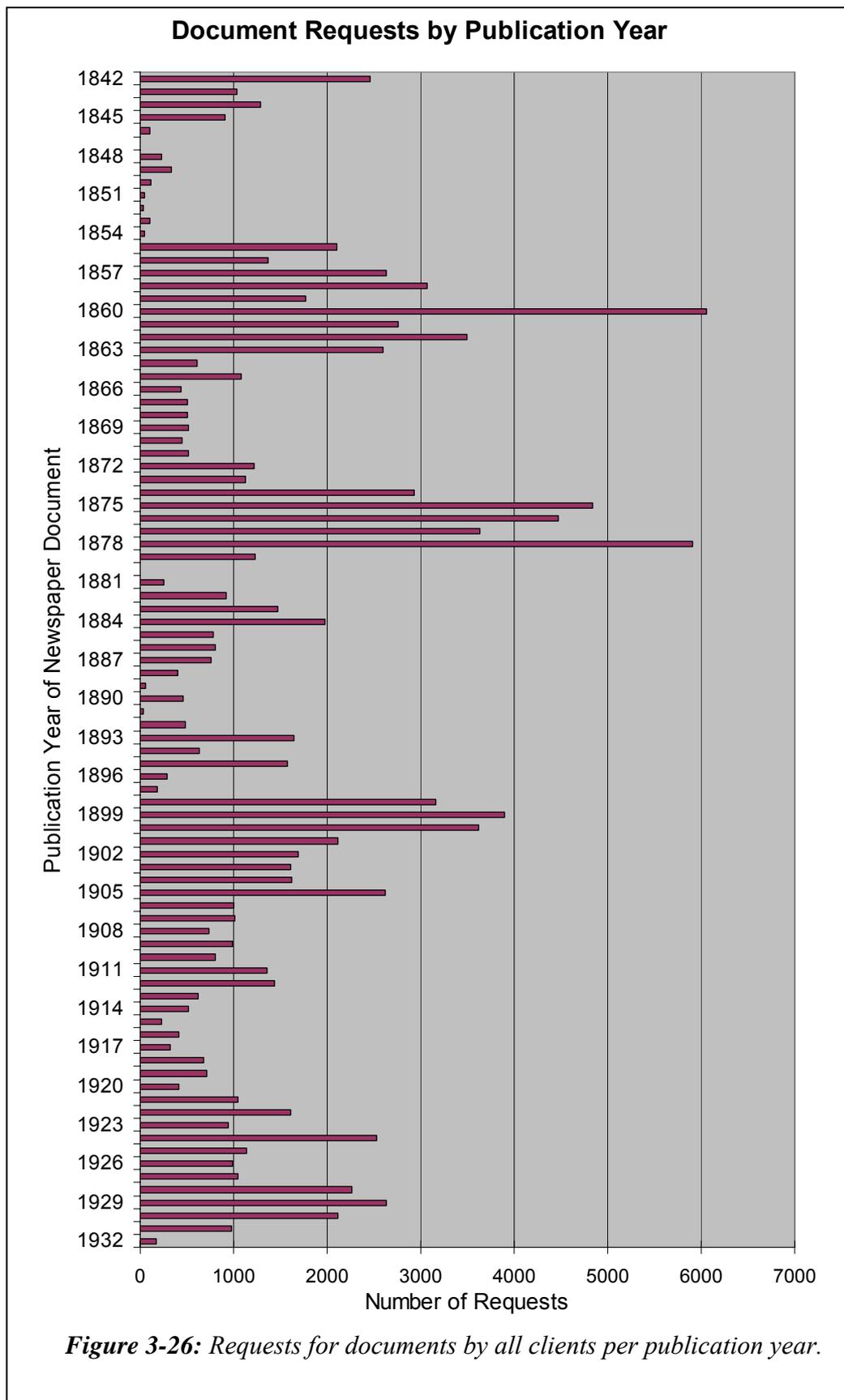
3.5.4 Which Publication Years Were Favoured?

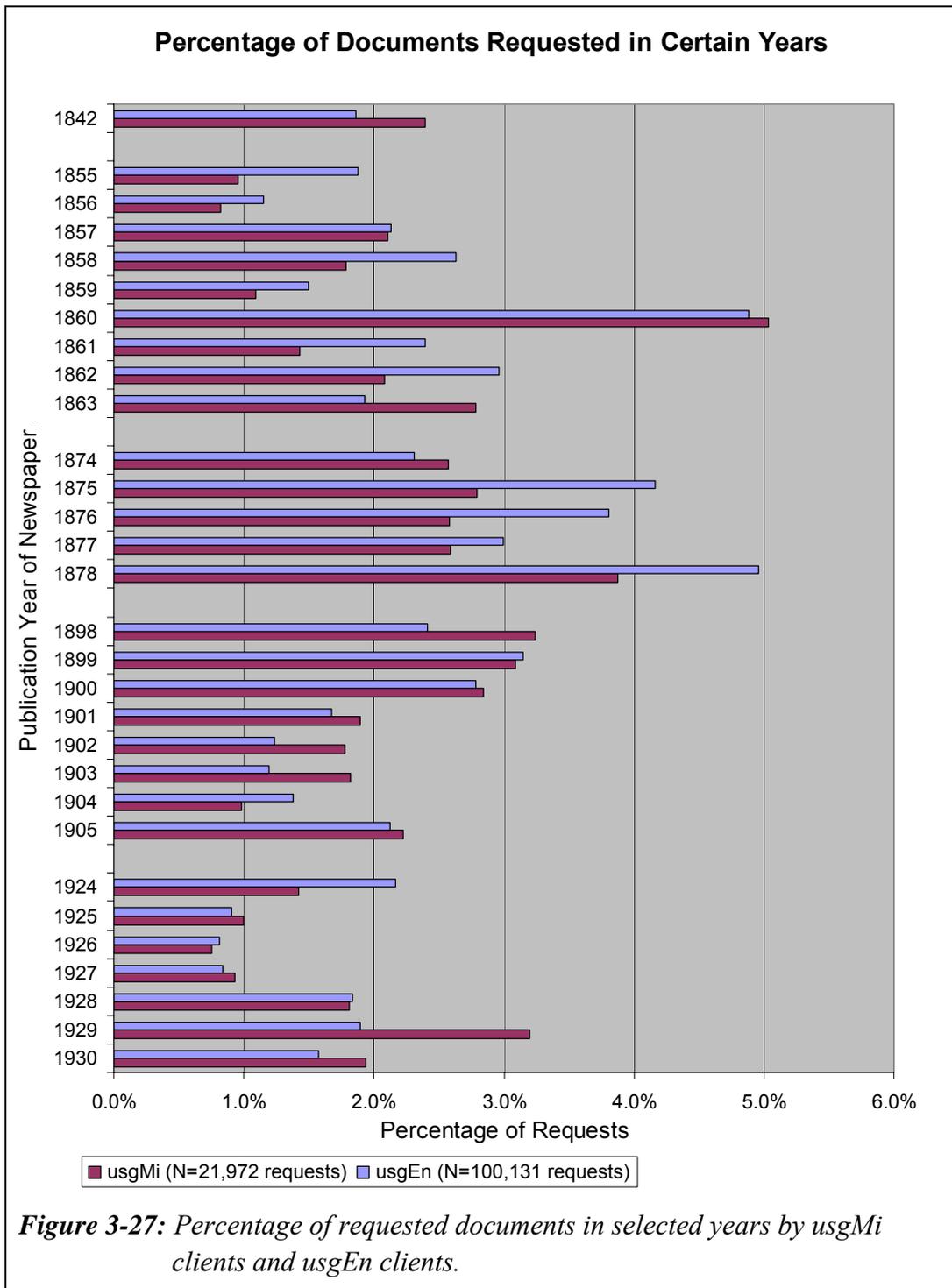
The periodicals in the Niupepa collection were published over a 90 year period, from 1842 to 1932. The count of pages requested from the Niupepa collection per year is displayed in Figure 3-26. This figure shows that publications in certain

years were requested more often by Niupepa clients. The most favoured years were 1842, 1855–1863, 1874–1878, 1898–1905 and 1924–1930.

The popularity of some of these years can be explained. It is not unexpected to see the first year of publications (1842) popular as this is the time period researchers and scholars access to see the first examples of early writing in te reo Māori. There are also English Abstracts available for these early years, making the content understandable to clients with limited or no reo Māori knowledge. The next two time periods, 1855–1863 and 1874–1878 are time periods where bilingual newspapers were published and also English Abstracts are available for those newspapers published only in te reo Māori. Further research is needed to determine why the 1898–1905 and 1924–1930 time periods are so popular.

The most popular dates for which documents are requested have been listed again but this time, rather than the count of the requests, the percentages of the requests are listed. This listing is to determine which dates are important to usgMi clients and also which dates are comparatively important to usgEn clients. Because of the low count of usgBi requests, document dates requested by usgBi clients were not graphed. The usgMi and usgEn request percentages are displayed in *Figure 3-27* which shows that the publication years identified in *Figure 3-26* as important years for clients of the Niupepa collection are also important years for clients who use the Niupepa Collections in te reo Māori. While some small differences are noted in the years usgMi and usgEn clients request documents (for example 1875 and 1921) these differences are not statistically significant.





3.5.5 Documents With Associated Extracted Text Files

If Niupepa documents had associated text files were they requested more often? One would assume this to be true as not only would the documents be

discoverable by browsing series list and by browsing dates, they would also be discoverable by submitting a query. In the Niupepa digital library 93.1% of the documents had an associated extracted text file. Overall there were 118,557 requests (98.6%) for documents with associated text files and 1660 requests (1.4%) for documents without associated text files. On average there were 3.5 requests for every document with a text file and 0.9 requests for every document without a text file. Thus, Niupepa documents with an associated text file were requested 3.8 times more often than Niupepa documents without an associated extracted text file.

This was true for all clients of the Niupepa digital library no matter what language the digital library was used in. A similar analysis was undertaken specifically for usgMi clients and usgEn clients and the results are displayed in *Table 3-10*. The analysis was not undertaken for usgBi sessions as the request count was too low to give meaningful results.

Requests for Documents with Extracted Text Files							
Total Documents		UsgMi Requests			UsgEn Requests		
	Count	Count	%	Req/doc	Count	%	Req/doc
With Text	33,674	20,877	97.6%	0.62	95,609	98.8%	2.8
Without Text	1808	517	2.4%	0.29	1116	1.2%	0.6

Table 3-10: Comparing the number of requests for documents with and without associated extracted text files in usgMi and usgEn usage sessions.

As shown in *Table 3-10* there were 20,877 usgMi requests (97.6%) for documents with text files and 517 usgMi requests (2.4%) for documents without text files. On average there were 0.62 usgMi requests for every document with a text file and 0.29 usgMi requests for every document without a text file, implying that usgMi clients are 2.2 times more likely to request a document with an associated extracted text file.

There were 95,609 usgEn requests (98.8%) for documents with text files and 1116 usgEn requests (1.2%) for documents without text files. On average there

were 2.8 usgEn requests for every document with a text file and 0.6 usgEn requests for every document without a text file, implying that usgEn clients are 4.6 times more likely to request a document with an associated extracted text file.

It is expected that documents with associated extracted text files would be requested more often than documents without. There are more ways in which to request these documents and more forms of the document to request. It would be expected, however, that both usgMi and usgEn would request these documents at an equally high rate. This was not the case. The figures suggest that usgMi clients were less than half as likely to prefer a document with an associated extracted text file.

3.5.6 Documents With Associated English Abstracts

Analysis was undertaken to determine if newspaper pages that had associated English Abstracts were requested proportionally more often than newspaper pages without. One would assume this statement to be true as not only would the documents be discoverable by browsing and searching they would also be discoverable by requesting an English Abstract and following the links from the abstract to the document. There were 22.7% of the Niupepa documents with an associated English Abstract.

There were 0.60 usgMi requests for every document with an English Abstract and also 0.60 usgMi requests for every document without an English Abstract file. There were 2.9 usgEn requests for every document with an English Abstract and 2.7 usgEn requests for every document without an English Abstract file. The usgBi data was not large enough to provide conclusive results.

The above figures suggest that having associated English Abstracts with documents did not affect the number of requests for those documents by usgMi clients. It did affect the number of requests for those documents by usgEn clients. UsgEn clients were 1.1 times more likely to request a document if it had an associated English Abstract.

3.5.7 How Long to Read Documents?

Reading of Niupepa documents has been defined as apparent document reading (see §2.5.1). An apparent read time of a document is the time interval from when one document is requested to when another document or page of the interface is requested. We cannot be sure that the client was only reading the document during this time; hence, these times are coined apparent read times.

The average apparent document read times by usgMi clients of the different document types were calculated and are displayed in *Table 3-11*. Also displayed are the average apparent read times by usgEn clients, by usgBi clients and the total average apparent read times.

Average Apparent Read Times of Documents				
(all times in seconds)	usgMi	usgEn	usgBi	All
Average Read of all documents	71.5	55.3	71.7	58.3
Average Read of Full size image	112.0	93.7	120.0	96.9
Average Read of Extracted Text	53.0	32.6	47.8	36.7
Average Read of Preview Image	35.5	26.8	23.9	27.9

Table 3-11: Comparing average apparent read times of document types in usgMi, usgEn and usgBi usage sessions.

The average apparent read times listed in *Table 3-11* suggest that usgMi clients spend, on average, less than two minutes reading the full size images, less than one minute reading the Extracted Text formats and a little over half a minute reading the preview images.

Table 3-11 indicates that all three types of usage clients spend the longest times reading the full size images and the least time reading the Preview Images. The figures also show that the usgMi sessions have longer apparent read times than the usgEn sessions for all types of documents.

3.5.8 Summary—Document Access Behaviours

This section analysed how te reo Māori clients in usage sessions accessed Niupepa documents. Statistics drawn from TLA examined seven document retrieval behaviours exhibited by usgMi clients.

The statistics revealed that using a query was the most favoured method used by usgMi clients to access documents. Using a query was six times more popular than the next most favoured method, browsing the collection by publication listing. Requesting documents in their extracted text form was twice as popular with usgMi clients as requesting documents in their full size image form. UsgMi clients showed a strong preference for accessing documents written in te reo Māori and were the clients least likely to access documents that weren't written in te reo Māori. There were a higher number of requests by usgMi clients for documents that were published in the years: 1842, 1857–1863, 1874–1878, 1898–1905 and 1924–1930. UsgMi clients were more than twice as likely to request a newspaper page if that page happened to have an associated extracted text file. UsgMi clients did not show an increased preference for newspaper pages with associated English Abstracts files. UsgMi clients, on average, required almost two minutes to read a full size image, almost one minute to read an extracted text file and 35 seconds to read a newspaper preview image.

This brief summary concludes the examination into how Niupepa documents were retrieved by clients in te reo Māori. The analysis of how the Niupepa digital library is used by clients in te reo Māori is undertaken in three parts; an analysis of how the interface was used, of how documents were accessed and on how queries were undertaken. Having completed the analysis of how documents were accessed, the next stage is to examine how queries were submitted by usgMi clients. This analysis begins in the following section.

3.6 Query Behaviours

The purpose of this section is to analyse query behaviour of usgMi clients in the Niupepa digital library. Query is the term used to describe the action taken by a client who has entered a query term or terms into the input box of the Niupepa search engine and then clicked the ‘Rapua/Begin Search’ button.

Query statistics were generated by running Perl scripts over the Niupepa web log dataset. There are nine query behaviours analysed in this section:

- the number of queries submitted in usgMi sessions
- the number of query terms used in usgMi queries
- the topics queried by usgMi clients
- the language of query term
- the altering of query parameters in usgMi queries
- the use of query operators in usgMi queries
- the number of documents requested following a usgMi query
- the formats favoured following a usgMi query
- the apparent document read times following a usgMi query.

These nine query behaviours comprise the next nine sub-sections and provide a description of how te reo Māori clients queried the Niupepa digital library.

3.6.1 How Many Queries Were Submitted?

The numbers of queries submitted in each usgMi session were analysed by TLA. There were 975 (57.3%) usgMi sessions that involved at least one query. In comparison, there were 3,317 (67.0%) usgEn sessions and 72 (80.0%) of usgBi sessions that also involved the submission of at least one query. The percentages of queries submitted per usgMi sessions are displayed in *Figure 3-28*. Also displayed are the percentages of queries per usgEn and usgBi sessions.

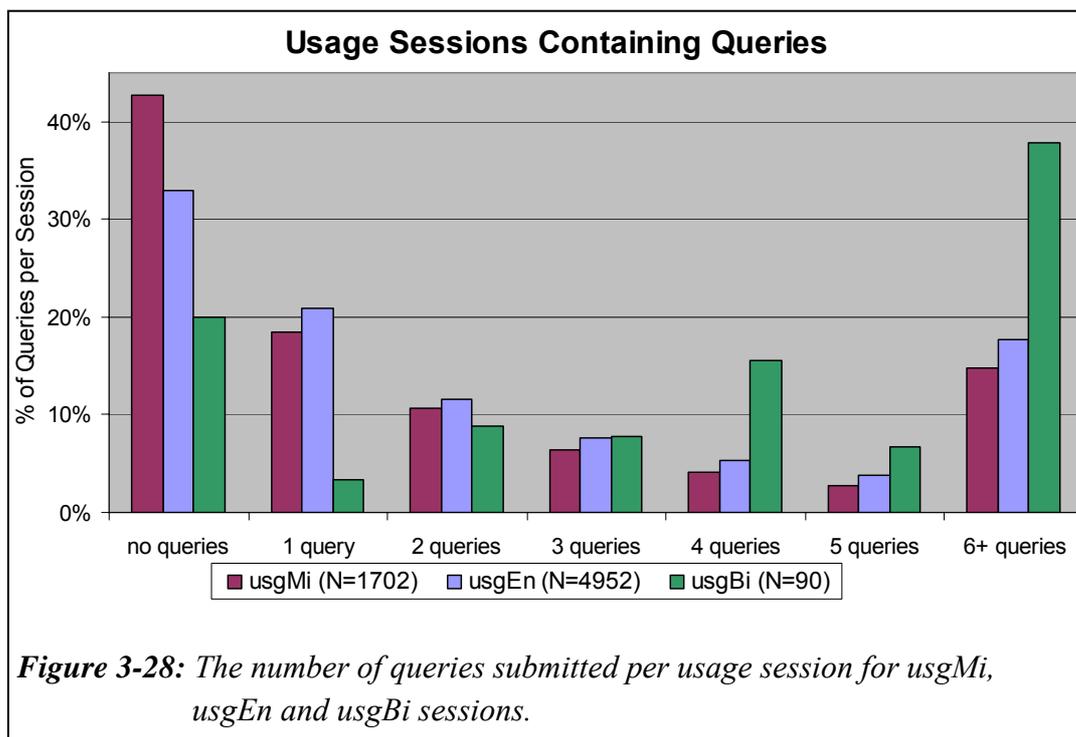


Figure 3-28 shows the usgMi sessions have a stronger presence in the nil to low number of queries submitted while the usgBi sessions have a stronger presence in the medium to high number of queries submitted. The usgEn sessions fall between these two extremes.

It was reported earlier (in §3.5.1) that 77.7% of documents requested by usgMi clients were a result of undertaking a search, which seems to contradict the numbers and figures recorded above. What these numbers do indicate is that usage session queries led to a high number of documents being accessed. The number of documents requested following a query is discussed further in §3.6.7.

3.6.2 How Many Query Terms Were Submitted?

The number of query terms submitted with each usgMi query was examined. Almost half of all usgMi queries consisted of just one query term; just over a quarter consisted of two query terms and the other quarter of usgMi queries

contained three or more query terms. The actual numbers and percentages of query terms per query are listed in *Table 3-12*.

Number of Query Terms Submitted Per Query								
	usgMi		usgEn		usgBi		Total	
Total Queries	4,767	100%	17,227	100%	504	100%	22502	100%
1 term/query	2233	46.8%	8554	49.6%	237	47.0%	11024	49.0%
2 terms/query	1412	29.6%	5475	31.8%	166	32.9%	7053	31.3%
3+ terms/query	1122	23.5%	3202	18.6%	101	20.0%	4425	19.7%

Table 3-12: The number of query terms submitted per query in usgMi, usgEn, usgBi and the total usage sessions.

While the percentages of query terms per query look similar for queries undertaken by usgMi, usgEn and usgBi clients the numbers are in fact significantly different. A chi square analysis on the counts generates a p value of $2.5 * 10^{-13}$ for the usgMi-usgEn comparison and a p value of $9.7 * 10^{-11}$ for the usgMi-usgBi comparison. These figures indicate that the usgMi clients had a statistically significant different preference for the number of terms that were submitted per query than the usgEn and usgBi clients. It can be seen that the usgMi clients submitted the highest number of queries with three or more query terms (23.5%). The usgEn clients submitted the highest number of queries with just the one query term (49.6%). The usgBi clients, by contrast, have the highest percentage (32.9%) for submitting queries which include two query terms.

3.6.3 What Topics Were Queried?

All query terms submitted by clients of the Niupepa digital library were recorded in the web log files. The TLA process collated these terms and grouped them into usage session types. Clients in usgMi sessions submitted 3,379 query terms in 4,766 queries. The term *tioriori* (a figurative term for a warrior) was entered the most often (26 queries) by usgMi clients.

Clients in usgEn sessions submitted 10,797 query terms in 17,227 queries. The term *whakatauki* (proverb) was entered the most often (62 queries) by usgEn clients. Clients in usgBi sessions submitted 322 query terms in 504 queries. The term *arawa* (a tribal name) was entered the most often (13 queries) by usgBi clients.

The top 20 requested query terms by usgMi clients, by usgEn clients and by usgBi clients are listed in *Table 3-13*. Similar terms with different capitalisation or different use of quotation marks are treated as different terms in this list.

The query terms were grouped into topic categories. The fifteen topic categories defined to classify the query terms:

- Names of people—specifically individual people, e.g. *Rewi Maniapoto*
- Iwi (tribes)—includes hapū (sub tribe), nationality and ethnicity e.g. *Tuwharetoa, Chinese*
- Social institutions and roles—e.g. *Kingitanga, kaitiaki*
- Place names—e.g. *Kawhia, Parihaka*
- Historic events and phenomena—e.g. *Treaty of Waitangi, Waikato Wars*
- Cultural activities and concepts—e.g. *haka, whaikorero*
- Cultural phenomena—items of an ethereal nature e.g. *taniwha,*
- Actions and attributes—e.g. *aroha, koingo*
- Geographical features—e.g. *moana, oneone*
- Celestial and seasonal phenomena—e.g. *matariki, christmas*
- Publication and information sources—e.g. *te hokioi, obituaries*
- Plant and animal names—e.g. *apple, tohora*
- Objects—general inanimate objects e.g. *pounamu, rorohiko*
- Language—greetings and particles—e.g. *e hoa, taku*
- Ambiguous—words that cannot be categorised because they have possible referents in several categories e.g. *Te Paki o Matariki* (is this a celestial object, social institution or a publication?)

Query Terms Submitted Most Often		
usgMi	usgEn	usgBi
tioriori (26)	whakatauki (62)	arawa (13)
friend (24)	makaka (53)	parihaka (8)
harakeke (22)	Whakaki (48)	pataua (8)
Tuwhenua (19)	waiata (47)	Ratana (8)
nakahi (14)	awaitaia (40)	"he kupu whakamarama" (5)
rangatiratanga (13)	haka (39)	kingitanga (5)
tuhoe (11)	matariki (32)	ngapuhi whakatauki (5)
mangamaunu (10)	Te wharepu (28)	taipari (5)
Kingitanga (9)	parihaka (27)	waka hokohoko (5)
tutakangahau (9)	treaty (27)	"local government" (4)
pounamu (8)	oruanui (26)	hawaii (4)
rorohiko (8)	motakotako (24)	he kupu whakamarama (4)
taharoa (8)	Tapiri (24)	king (4)
tawhiao (8)	Parihaka (23)	ko au ko ngati whatua (4)
whiro tirea (8)	treaty of waitangi (23)	land disputes (4)
1900 (7)	utiku potaka (23)	ngati koroki ki raukawa (4)
moana (7)	Chinese (20)	proverb (4)
rangatira (7)	pipiwharau (19)	tohora (4)
taku (7)	ihaka paehoka (18)	turanga baker (4)
Te Pipiwharau (7)	kewene (18)	waikato wars (4)
waewae kura (7)	Mu+aupoko+ (18)	wao tapu nui a tane (4)
Whakaminenga 1835 (7)	Tainguru (18)	
	tawhiao (18)	
	te puke ki hikurangi (18)	

Table 3-13: The query terms (with counts) most commonly submitted by usgMi, usgEn and usgBi clients.

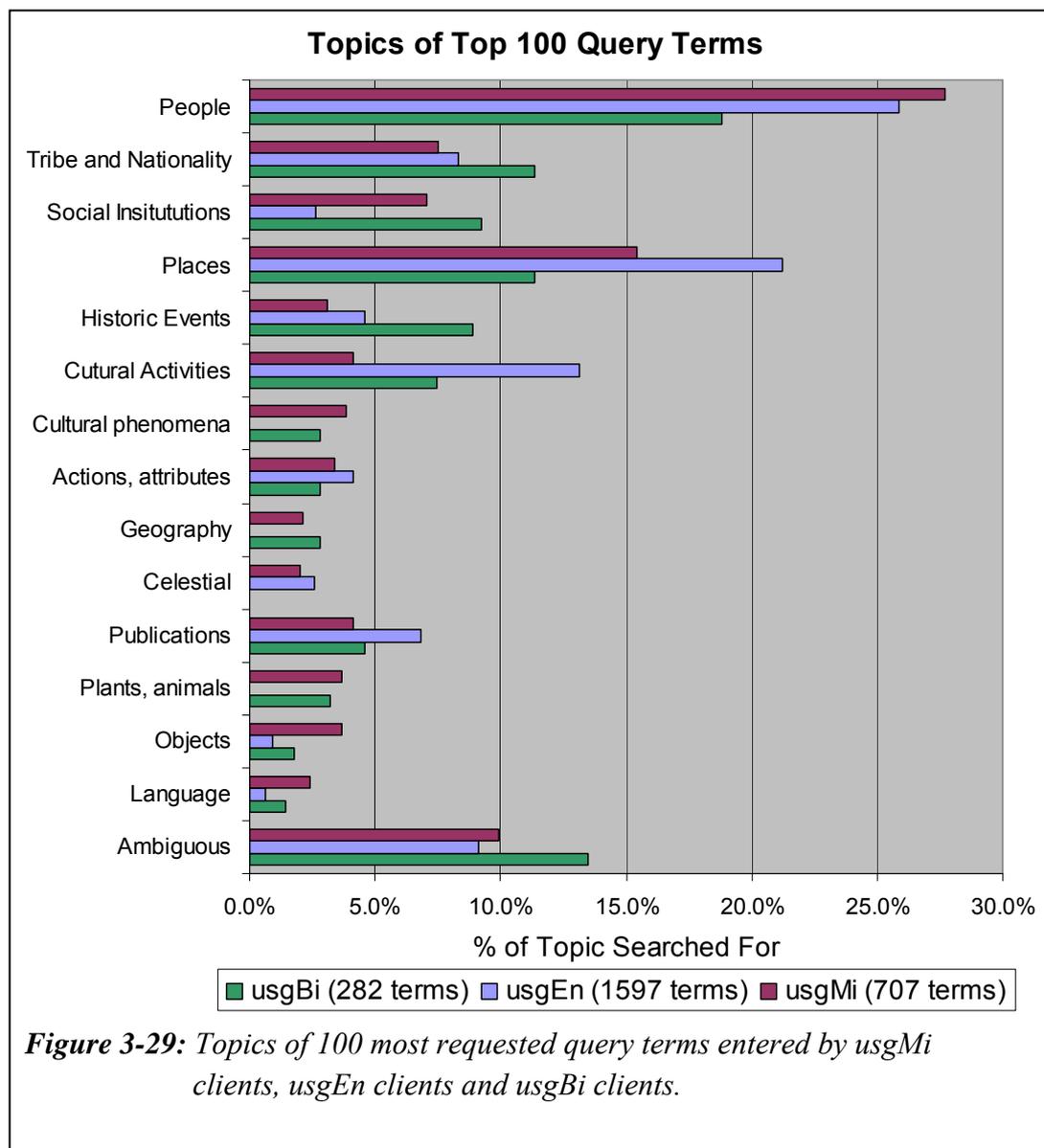
Trying to determine the topic of a query by simply examining the query term is highly speculative; there is no definite way to tell the purpose of a user's searching activity simply by noting the query term that has been submitted.

For example, the most common term submitted by usgMi clients was *tioriori*. There were 26 requests for this term (the next highest was 24). Williams (1985) gives three definitions of the base word, *tiori*: 1) hold up to view, 2) wave to and fro, 3) conspicuous. Williams goes further to explain that *tioriori* is a form of kite, sometime called *manu tioriori*, which is applied to a decoy bird and figuratively to a brave warrior. As the query term has no capitals one could expect that this query term could perhaps be classified as either an attribute or an object. However, when this term is searched in the Niupepa collection 16 documents are found that include this term, 13 of which are names of people and the remaining three are from the phrase *manu tioriori*. One could speculate that some of these results must be useful to the user, hence the high number of searches for this term. Further, the term *tioriori* appears in other queries entered by usgMi clients: *ihaia tioriori* (four requests), *niwha wharerangi tioriori* (three requests), *tioriori kaiwhakawa* (three requests), and *tioriori te haira* (one request), all are queries for an individual. Consequently the term *tioriori* was categorised, albeit somewhat apprehensively, as a query for a person.

Defining the query terms into categories is a laborious process that involves a lot of speculation. Consequently, only the top 100 unique query terms were examined and categorised into the topic headings. For the usgMi sessions this categorisation involved 111 terms that were queried 707 times. For the usgEn sessions this categorisation involved 102 terms that were queried 1,596 times. For the usgBi sessions it involved 100 terms that were queried 282 times. These query terms were categorised into topic categories, the percentage of each term per category was graphed and is displayed in *Figure 3-29*. Although this method of analysis is speculative some general query behaviours are apparent.

The most common topics queried by usgMi clients were names of people and place names. Although 10% of the topic categories were ambiguous usgMi clients did show an interest in all the other topic categories.

Names of people and place names were also the most popular query categories of usgEn and usgBi clients. UsgEn clients also showed a high interest in cultural activities and concepts, publication and information sources and the topic tribes, nationalities and ethnicities. UsgBi clients showed a high number of ambiguous query topics and interests in a wider range of topics.



3.6.4 What Was the Language Of the Query Terms?

The query terms were written in either te reo Māori or English. The content of the Niupepa digital library that was queried consisted not only of the 35,482 newspaper pages but also 313 English Abstracts and 22 Bibliographic Commentaries. Of this content, 53.3% was available only in te reo Māori, 43.9% was available in te reo Māori and English and the remaining 2.8% was available only in English. As there was a large amount of content in te reo Māori, query terms submitted in te reo Māori are likely to yield more results.

The language of the query terms was noted for the top 100 query terms in each usage category and the percentages have been plotted in *Figure 3-30*.

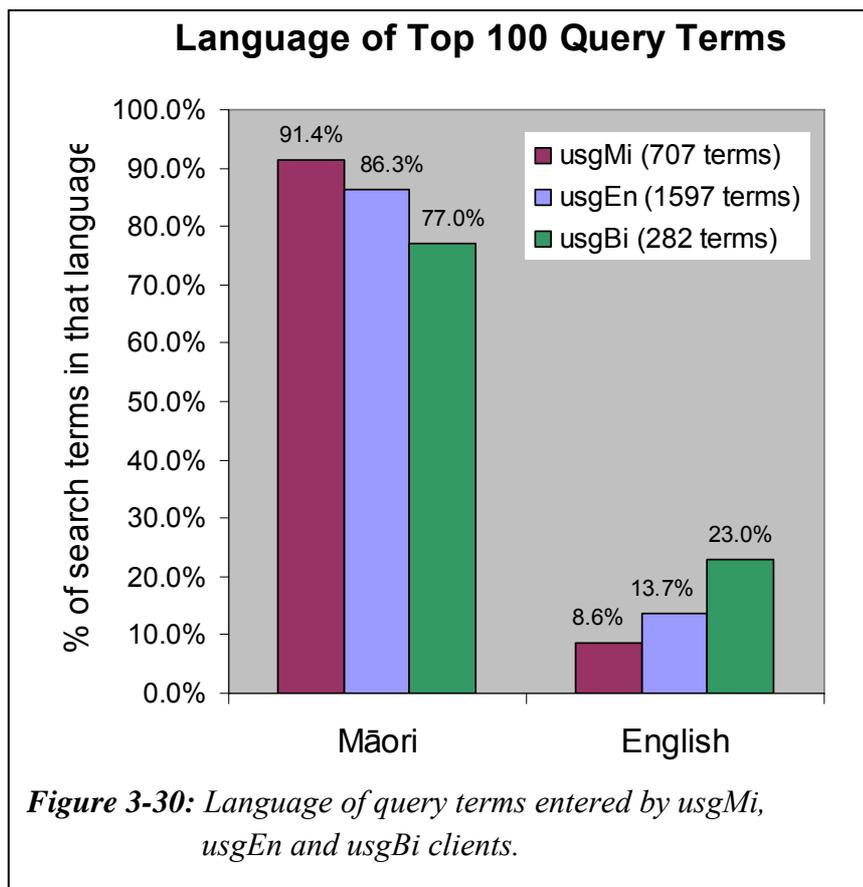


Figure 3-30 shows that 91.4% of the top 100 query terms submitted by usgMi clients were written in te reo Māori. The query terms and their frequencies entered

in English by usgMi clients were: *friend* (24), *maori land court* (6), *death notices* (5), *east coast watch* (5), *Troy* (4), *oil* (4), *prosser* (4) and *Chinese* (4).

The usgEn sessions also had a high percentage of query terms in te reo Māori at 86.3%. This high percentage perhaps suggests that even though these clients are undertaking activities with the interface language set to English they do have an understanding of te reo Māori words. UsgBi sessions have the lowest percentage of terms submitted in te reo Māori at only 77.0%. This lower percentage could suggest that clients defined as usgBi clients were specifically targeting English query terms to assist in the information retrieval process.

3.6.5 Were Query Parameters Altered?

The Niupepa web site offers the user the ability to alter a number of query preferences. These include options to match upper case and lower case characters, to ignore word endings, to retain a list of query terms used, to set the number of results per page and to search on *Some* or *All* of the query terms. Aside from the *Some/All* option, all of these query preferences must be altered on an interface page separate from the Query page, that is the Preferences page (see *Figure 3-2*).

Users seldom altered the query parameters on the Preferences page. Less than 0.13% of usgMi clients altered query parameters on the Preferences page. This behaviour was also displayed by usgEn clients, less than 0.22% altering these preferences, and usgBi clients, less than 0.37% altering these preferences.

However, the one query parameter that was able to be altered on the actual Query page, the *Some/All* option was altered by 10.6 % of the Niupepa digital library clients. The *Some/All* option was least altered by usgMi clients; only 8.4% of the usgMi clients altered this parameter. However, 10.7% of the usgEn clients and a high 23.3% of the usgBi clients chose to alter this parameter when undertaking queries.

3.6.6 Were Query Operators Used?

Operators can be used in queries to broaden or narrow search results. There are two types of operators recognised by the Niupepa digital library search engine; logical operators (and, or, not, *, +, -), and phrase defining operators (quotation marks).

Results from TLA showed that the use of logical operators was low when undertaking a query. Less than 0.9% of usgMi queries used logical operators. The use of logical operators in usgEn queries (0.4%) and usgBi queries was also low (0.5%).

However, there was significant use of the quotation marks to define precise phrase searching. Of the 4,767 usgMi queries submitted 825 (17.3%) involved the use of quotation marks to specify exact phrase matching. This rate was higher than in usgEn queries (6.4%) and usgBi queries (8.3%). It has been noted that usgMi clients entered more query terms when undertaking queries (see §3.6.2); thus it seems probable that usgMi clients would also be more likely to use an operator, such as quotation marks, to assist in defining results in multiple term queries.

3.6.7 Documents Requested Following a Query

The number of documents requested by usgMi clients following the submission of a query was revealed by TLA. There were 4,767 queries submitted in usgMi sessions, directly leading to the request of 16,081 documents; a return rate of 3.4 documents per query. There was also an average of 4.6 documents requested by usgEn clients following a usgEn query and 3.2 documents requested by usgBi clients following a usgBi query.

These figures indicate that usgMi clients (and usgBi clients) did not examine as many search results as usgEn clients. This low request count following a query could be because usgMi (and usgBi) searches were more accurate in returning what the client was looking for or it could simply average that usgMi (and usgBi) clients did not like browsing through the search results pages.

3.6.8 Formats Favoured Following a Query

Statistics produced from the TLA process showed which document formats were requested following the submission of a query. The figures were grouped by language of use, allowing the collation of numbers and percentages of documents requested by usgMi clients following a usgMi query. These percentages are displayed in *Table 3-14*, along with percentages for usgEn and usgBi clients. To assist with a query/non-query comparison, figures and percentages are also listed displaying the document formats requested from activities that did not include a query.

Document Formats Requested following a Query								
	usgMi		usgEn		usgBi		Total	
	Non Query	After Query	Non Query	After Query	Non Query	After Query	Non Query	After Query
Total	6,167	16,081	22,848	79,192	683	1,589	29,698	96,862
Extract. Text	40.9%	69.7%	48.3%	56.8%	51.5%	55.8%	46.8%	59.0%
Full image	50.3%	26.7%	38.0%	38.4%	41.9%	33.4%	40.6%	36.3%
Preview	7.2%	3.5%	12.1%	4.6%	6.1%	10.6%	10.9%	4.5%
Undefined	1.7%	0.2%	1.7%	0.2%	0.4%	0.3%	1.6%	0.2%

Table 3-14: Comparing document formats requested not involving a query with those requested following a query from usgMi, usgEn and usgBi clients.

Table 3-14 reveals that the usgMi clients were much more likely to request an extracted text file after a query than they were if a query was not involved. This characteristic was shown by all client types. Perhaps this increased preference is understandable as the icon links on Niupepa results page all link to the Extracted Text version of the Niupepa documents. However, the increased preference for an extracted text format is much higher in usgMi sessions (28.8%) than in usgEn sessions (8.5%) and in usgBi sessions (4.3%).

3.6.9 Document Read Times Following Query

The apparent read times of documents following a query were compared with the apparent read times of documents when no query was involved. These times, in seconds, are displayed in *Table 3-15*.

(times in seconds)	usgMi		usgEn		usgBi	
	Non Query	After Query	Non Query	After Query	Non Query	After Query
All Documents	87.3	65.8	57.7	54.6	89.2	63.4
Full size Image	126.8	101.8	85.4	96.0	128.1	115.6
Extracted Text	51.4	53.4	44.9	29.7	66.3	40.5
Preview Image	27.1	43.7	27.4	26.3	20.1	25.0

Table 3-15: Comparing apparent read times of non-query documents with apparent read times of query documents by usgMi, usgEn and usgBi clients.

It appears that the usgMi apparent read times of documents were affected by whether or not the document had been requested by a query activity. In particular, the usgMi apparent read times of a full size image were reduced following a query (126.8 to 101.8 seconds) and the usgMi apparent read times of preview images were increased (27.1 to 43.7 seconds) following a query. These times suggest that usgMi clients spent less time reading a full size image after submitting a query and more time reading a Preview Image. Read times of extracted texts remained similar whether a query was submitted or not.

Table 3-15 shows that the usgEn and usgBi apparent read times of extracted texts did change significantly following a query, and smaller changes were also noted in the usgEn and usgBi apparent reads of full size and preview images.

3.6.10 Summary—Query Behaviours

Section 3.6 analysed how te reo Māori clients in usage sessions undertook queries in the Niupepa digital library. Statistics drawn from TLA examined nine query behaviours exhibited by usgMi clients.

The statistics revealed that over 57% of usgMi sessions did involve the submission of a query. Almost half of all usgMi queries consisted of just one query term; just over a quarter consisted of two query terms and the other quarter involved three or more query terms. The most common topics queried by usgMi clients were names of people and places with a small interest also shown in a wide range of other topics. Of the top 100 query terms submitted by usgMi clients, 93% were written in te reo Māori. Query parameters were not altered on the Preferences page (0.1% of usgMi clients) but the Some/All parameter was altered on the Query page by 11% of the usgMi clients. Logical operators were rarely used by usgMi clients (0.9%) but quotation marks were used by 13% to undertake exact phrase matching. For every query submitted, usgMi clients accessed 3.5 documents and were 30% more likely to request the Extracted Text format following a query than they were if a query had not been submitted. After a query was submitted the apparent read times reduced for full size image formats, increased for preview image formats and remained approximately the same for Extracted Text formats.

Having summarised how te reo Māori clients used the Niupepa digital library we are now in a position to deal with two further usage questions: how did usage in te reo Māori differ from usage in English, and how was the Niupepa digital library used by those clients who chose to use both te reo Māori and English? These questions will be addressed in the next two sections.

3.7 Te Reo Māori Usage/English Language Usage

Although the purpose of this chapter is to analyse how te reo Māori was used by clients of the Niupepa digital library, there is also value in detailing how the usage in te reo Māori differed from the usage in English. If differences can be defined then these differences can be incorporated into digital library design and implementation, which could further increase the assistance that digital libraries can give to indigenous languages. The language usage differences have been subdivided into; behaviour differences in expMi and expEn sessions and behaviour differences in usgMi and usgEn sessions.

Behaviour differences in exploratory sessions

Even though the exploratory sessions represented less than 1% of the requests to the Niupepa collection they did highlight three important behaviour differences. First, the expMi sessions were, on average, longer in time (199 versus 136 seconds) and in the number of requests (2.4 versus 2.3 requests) than expEn sessions.

The second behaviour difference noted was that in exploratory sessions different interface pages were requested by expMi and expEn clients. Both usage session types requested the home page as a first preference. However, as a second preference the expMi clients preferred to request the About English Abstracts page, whereas the expEn clients preferred to access the NZDL home page.

The average apparent read time of interface pages was longer for clients who used the interface in te reo Māori. The average expMi apparent read time of interface pages was 2.4 minutes compared with 1.8 minutes for expEn clients.

Behaviour differences in usage sessions

The Niupepa digital library was used more by usgEn clients than by usgMi clients. UsgEn clients were responsible for 79.6% of the total usage requests while usgMi clients were responsible for 18.1% (2.3% were submitted by usgBi clients). On average, usage sessions in English were longer than usage sessions in te reo Māori. The usgEn sessions were longer in time (16.8 compared with 21.3 minutes) than the usgMi sessions and in the average number of requests (18.4 compared with 27.8 requests) per session.

Similarly, the average apparent read time of *interface pages* by clients who used the Niupepa digital library in te reo Māori was longer than for those who used it in English. The average usgMi apparent read time of interface pages was 52 seconds compared with 49 seconds for usgEn clients. A similar pattern occurred in average apparent read times of *documents*; the average usgMi apparent read time of documents was 72 seconds, the average usgEn apparent read time was 55 seconds.

The most popular method of accessing documents was by submitting a query. Queries were not as popular with usgMi clients (77.7%) as with usgEn clients (81.6%). The second most popular method to select documents with usgMi clients was to use the series list while the second most popular method with usgEn clients was to use the date list.

Clients of the Niupepa digital library in te reo Māori (usgMi) preferred to request documents that had a reo Māori component. UsgMi clients displayed a higher preference for requesting te reo Māori newspapers, documents that were written in te reo Māori. Alternatively, usgEn clients had higher preferences for requesting bilingual newspapers, English language newspapers, English Abstracts and Commentary documents (which were written primarily in English). UsgMi clients were twice as likely to access documents with an associated text file as they were to access files without an associated test file. UsgEn clients were almost five times as likely to access documents that had an associated extracted text file. UsgEn clients showed a higher preference for accessing documents with an associated English Abstract whereas usgMi users did not show preferences for documents with an English Abstract.

As mentioned above, submitting queries was less popular with usgMi clients. However, the usgMi clients showed a higher preference for submitting queries with three or more query terms, whereas the usgEn clients also showed a preference for submitting terms with only one or two query terms. Perhaps because of this difference the usgMi clients were three times as likely to use quotation marks surrounding the query terms to activate phrase searching. The usgMi clients were less likely than the usgEn clients to alter the Some/All parameter to further refine queries. Fewer documents were requested by usgMi clients (3.6) following a query than by usgEn clients (4.8).

The submission of a query affected the types of documents that were requested. When a query had not been submitted usgMi clients preferred to request full size images. However, following a query usgMi clients had a much higher preference for requesting extracted text documents, while the preferences for requesting full size images and preview images reduced by almost half. UsgEn clients did not show the same differences following a query. UsgEn clients displayed the highest preference for accessing an extracted text document whether a query was submitted or not. The only difference following a query by usgEn clients was a reduction in the preference for accessing the preview images, which subsequently led to an increase in the preference for accessing the extracted text documents.

The submission of a query also affected the apparent read times of documents. For usgMi sessions, the apparent read times following a query decreased for full size images, remained the same for extracted text documents but increased for preview images. For usgEn sessions, the apparent read times increased following a query, increased for full size images, decreased for extracted text documents and stayed the same for preview images.

A reo Māori to English interface language switch is something that usgEn showed a strong preference for. When usgEn clients undertook this switch it usually occurred at the start of a session. When usgMi clients switched the interface language from reo Māori to English they did it towards the end of a session. Both user groups switched from English to te reo Māori, although not to the same extent and the switches were mostly occurred at the beginning or end of a session.

3.8 Bilingual Usage

The decision to divide the usage sessions into three types of language users, usgMi clients, usgEn clients and usgBi clients, was based on earlier research (T. T. Keegan & Cunningham, 2005a). This research suggested there were three distinct types of clients; those that undertook sessions primarily with the interface language set to te reo Māori (usgMi), those that undertook sessions primarily with the interface language set to English (usgEn), and those that undertook sessions with the interface language set to both te reo Māori and English for a significant portion of the session (usgBi).

Web log analysis undertaken in this chapter has suggested that these usgBi sessions were not the result of a separate and distinctive user group but, instead, were the result of a unique activity that was undertaken by a small number of clients. Only 1% of all usage sessions were classified as usgBi sessions. Of the 90 usgBi sessions, none was a return session, suggesting this is an activity that is undertaken only once by a client. At least 13 of the clients that undertook this activity also undertook activity as a usgEn client (11) or as a usgMi client (2).

The requirements for usgBi sessions ensured that these sessions would be long sessions. UsgBi sessions were defined as having at least 20% of the requests in both te reo Māori and English, at least ten requests and at least three language switches. Consequently, usgBi sessions were, on average, twice as long in duration as usgMi and usgEn sessions, and consisted of twice as many requests.

UsgBi clients requested the Preferences page almost twice as often as usgMi and usgEn clients. Aside from the home page, the Preferences page is the page on which the language of the interface is switched so it seems likely that usgBi clients were accessing this page more, as all usgBi sessions had at least three language switches. The average apparent reading time of interface pages was longer for usgBi sessions, again, an attribute possibly caused by the longer time it takes to switch the interface language through the Preferences page.

Three query activities were undertaken differently by usgBi clients; the number of queries submitted per session by usgBi clients was higher than by usgMi or usgEn clients, the number of English terms submitted per query was the highest in usgBi sessions, and the Some/All query parameter was altered by usgBi clients at least twice as often as by usgMi or usgEn clients. All three of these activities suggest that the usgBi sessions were sessions where submitting a query was a primary activity, and that both te reo Māori and the English language were used together in a creative way to assist with this querying activity. The usgBi clients accessed fewer documents per query than usgMi or usgEn clients, suggesting that perhaps these queries were readily returning the desired result.

Many of the other usage behaviours exhibited by usgBi clients were similar to those recorded for usgMi and usgEn clients. The apparent reading times of documents, the type of documents requested, the number of query terms, the topic of query terms, the activities following a query and the positions in the session where the interface language was switched were similar to, or between the extremes of, usgMi and usgEn behaviours.

These results suggest that te reo Māori, in conjunction with the English language, was used as a deliberate strategy to assist in information retrieval activities by bilingual clients. Bilingual clients have the ability to use the Niupepa digital library in te reo Māori or English. It appears that a small number of clients were choosing not to use it in one language or the other but, rather, were choosing to use it in *both* languages in the one session. This use of two languages appears to be an effective information retrieval strategy and was used by a small number of clients to request a relatively high number of documents.

3.9 Summary

The purpose of this chapter has been to use Transaction Log Analysis (TLA) to determine how te reo Māori was used by clients of the Niupepa digital library in 2005. This chapter began with a full description of the Niupepa interface screens and options available to clients.

TLA was used to filter and categorise all requests to the Niupepa digital library into one of three types of sessions; single request sessions, exploratory sessions, and usage sessions. By analysing this data it was possible to answer the question, was te reo Māori used by clients of the Niupepa digital library? Results from the analysis showed that te reo Māori was used by Niupepa clients in 20.5% of all requests. Te reo Māori was used in 35% of single request sessions, in 46% of the requests in exploratory sessions and in 19% of the requests in usage sessions.

Exploratory session analysis was undertaken to determine how clients used te reo Māori to explore the Niupepa web site but then not go on to use it. The 180 exploratory sessions only represented a small number of requests to the Niupepa digital library (less than 1%). These types of clients, expMi clients, only made two or three requests and mostly completed the session in less than half a minute. The Niupepa Home page was the interface page mostly requested by expMi clients, although the About English Abstracts page was also significantly requested.

As opposed to the limited amount of information available from exploratory session analysis, usage session analysis was able to provide more extensive information on how te reo Māori was used by Niupepa clients. Most of the requests (94%) to the Niupepa digital library were requests classified as usage session requests. Data on usage sessions showed that sessions undertaken by usgMi clients averaged 30 minutes and 18 requests. Only 6% of sessions were return sessions. The usage analysis was collated into three categories; interface usage by usgMi clients, document access by usgMi clients and queries submitted by usgMi clients.

Interface usage analysis by usgMi clients showed the Niupepa Home page was the interface page most requested, although the Preferences page was also significantly requested. The apparent read time of interface pages by usgMi clients was 52 seconds. The language of the interface was not switched in many usgMi sessions (only 8%); when it was switched it was likely to be from English to Māori and generally in the first half of the session.

Document access analysis by usgMi clients revealed using a query was the most favoured method used to access documents. Overall, requesting documents in the extracted text format was twice as popular with usgMi clients as requesting documents in the full size image format. UsgMi clients showed a high preference for documents written in te reo Māori and low preference for documents written in English. Documents with associated extracted text files were requested twice as much by usgMi clients. The average apparent read times by usgMi clients for documents were almost two minutes for a full size image, almost one minute for an extracted text file and 35 seconds for a preview image.

Analysis of queries submitted by usgMi clients showed that almost 43% of usgMi sessions did not involve the submission of a query. However when a query was submitted almost half consisted of just one query term, just over a quarter consisted of two query terms and the other quarter involved three or more query terms. Names of people and place names were the most common topics queried by usgMi clients. Most query terms (93%) were written in te reo Māori. Query parameters and logical operators were not used by usgMi clients when undertaking queries, aside from the Some/All parameter on the Query page and the speech quote operators to define phrase searching. On average usgMi clients accessed 3.5 documents following a query. Full size images were the document formats favoured by usgMi clients when a query had not been submitted; however post-query the extracted text format was favoured. Apparent read times of documents following a query increased for preview images, reduced for full size images and did not alter for Extracted Text formats.

Activities undertaken on the Niupepa digital library in te reo Māori were compared with activities undertaken in English. Differences noted in the statistics

showed that expMi sessions were longer than expEn sessions but usgMi sessions were shorter than usgEn sessions. Clients in te reo Māori appeared to take longer to read the interface pages and Niupepa documents. Querying was undertaken less by usgMi clients but when it was undertaken it often involved the use of more query terms and fewer results were examined following a query.

The final question raised in this chapter asked how te reo Māori had been used with the English language to assist in information retrieval. A small number of bilingual clients chose to use te reo Māori *and* English in the one usage session. These sessions, usgBi sessions, were longer sessions, concentrating on querying activities, and contained the highest mix of reo Māori/English query terms. UsgBi clients requested the lowest number of documents following a query, an average of 3.2; lower than usgMi clients who requested 3.4 documents and lower than usgEn clients who requested 4.6 documents following a query. These figures suggest that using te reo Māori with English was a strategy that led to more accurate query activity.

The information presented in this chapter has charted te reo Māori usage by clients of the Niupepa Digital library in 2005. It has distinctly shown how te reo Māori has been used by clients to undertake information retrieval behaviour in the Niupepa digital library. This reo Māori use has occurred without prompting or coercion, clients have chosen to use te reo Māori in a modern electronic medium.

The next question to be considered is whether the usage of the Niupepa digital library by reo Māori clients differs between those clients from Aotearoa and those clients who access the Niupepa digital library from overseas. The following chapter uses statistics generated from TLA to undertake a comparative analysis of international and national usage of the Niupepa digital library by clients who undertake their activities in te reo Māori.

4 Comparing International and National Reo Māori Usage

This chapter analyses how te reo Māori was used by international (non-Aotearoa) and national (Aotearoa) clients of the Niupepa digital library. Was te reo Māori used by international clients? If it was, was te reo Māori used in the same way by international clients as it was by national clients?

Statistics generated from Transaction Log Analysis (TLA) were used to compare te reo Māori usage by international and national Niupepa clients in 2005. A comparative analysis examines exploratory and usage activities of international and national clients. Usage activities examined include interface usage behaviours, document access behaviours and query behaviours. The international usgMi clients are shown to be distinct users with their own peculiar information retrieval behaviours.

The method used to differentiate between national and international requests has been described in §2.2.2. Two datasets were available because two mirror web servers served web requests to the Niupepa digital library in 2005. A server called Puka, located in Hamilton, Aotearoa, was responsible for serving all national requests. A server called Thor, in Lethbridge, Canada, was responsible for serving all international requests. Both servers independently logged the requests that they served into web log files. The Puka web file provided raw data about national requests. The Thor web file provided raw data about international requests.

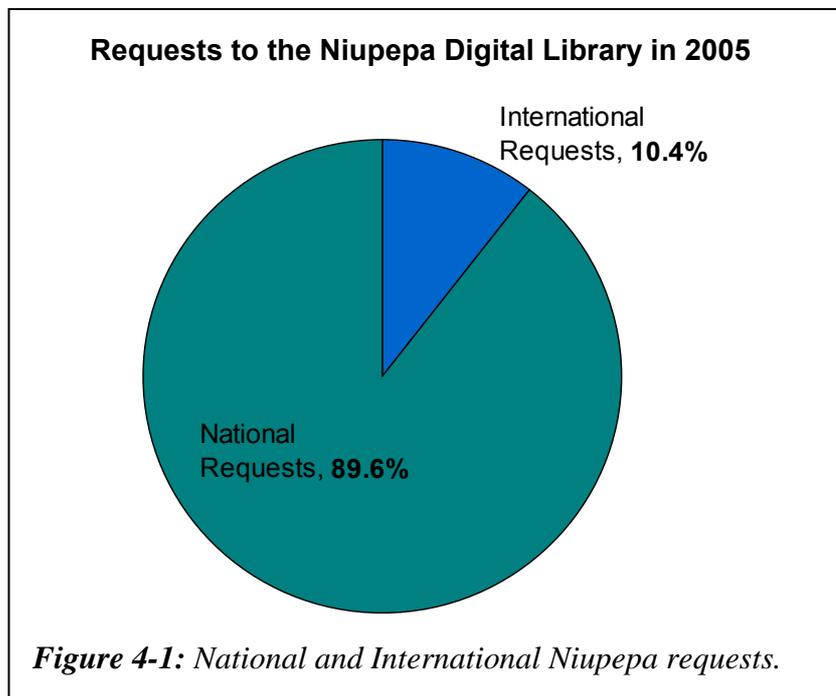
4.1 Was Te Reo Māori Used by International Clients?

This chapter details a comparative analysis of te reo Māori usage of the Niupepa digital library by national and international clients. However before this comparison can begin it must first be established that the Niupepa digital library was used by international clients and then it must be established that it was used by international clients in te reo Māori. This section uses statistics produced from TLA to answer these questions.

4.1.1 Did International Clients Use the Niupepa?

The Niupepa digital library is made available through the New Zealand Digital Library (NZDL) web site (see www.nzdl.org). The NZDL web site received a total of 1,133,543 raw requests in 2005 (see *Table 2-2*) of which 842,403 requests (74.3%) were logged as international requests and 291,140 requests (25.7%) were logged as national requests. However, most of the international requests were requests to other NZDL digital library collections, not the Niupepa digital library. There were only 28,501 international requests (3.4% of international requests) to the Niupepa digital library. In contrast, most of the national NZDL requests were to the Niupepa digital library and not to other NZDL digital libraries. There were 175,864 national requests (60.4% of national requests) to the Niupepa digital library.

Requests described as extraneous requests (see §2.3.2) were removed from the two sets of logs leaving 19,303 filtered international requests and 165,698 filtered national requests. These requests are displayed graphically in *Figure 4-1*.



Although the 19,303 international Niupepa requests are comparatively small in number (10.4%) compared with the number of national Niupepa requests there are still sufficient to undertake meaningful TLA.

When removing extraneous requests it was noted that the international requests had a much higher proportion of requests removed because of web robot activity. There were 9,122 (32.0%) web robot requests removed from the international Niupepa log. There were 8,055 (4.6%) web robot requests removed from the national Niupepa log. This indicates that although some web robot activity did originate in Aotearoa it was almost eight times more likely to have originated from outside Aotearoa.

International clients did use the Niupepa digital library to submit almost 20,000 requests.

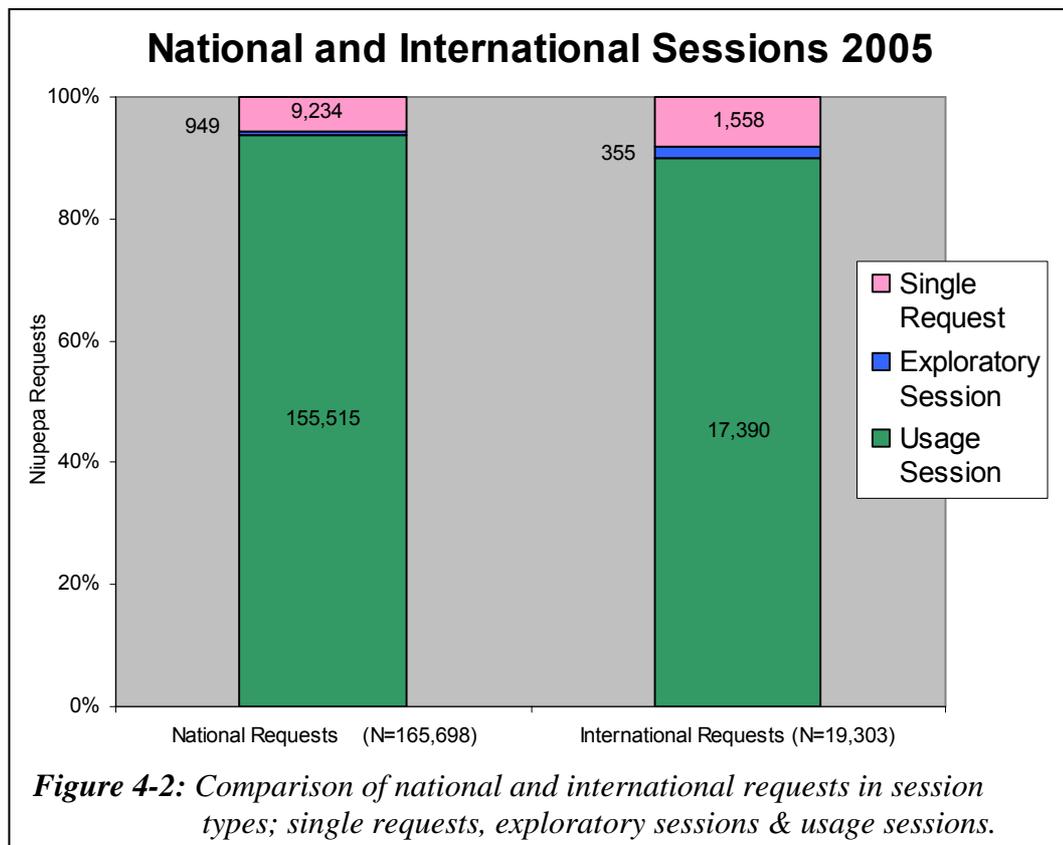
4.1.2 Did International Clients Show Interest in the Niupepa Digital Library?

The proportion of session types can give an indication of how interested prospective clients were in the Niupepa digital library. A high proportion of usage sessions combined with a low proportion of exploratory and single request sessions indicated a high level of interest.

TLA was used to separate the international and national requests into session types. The numbers and percentages of session types are displayed in *Figure 4-2*.

The 19,303 international requests consisted of 1,558 (8.1%) single requests, 355 (1.8%) exploratory session requests and 17,390 (90.1%) usage session requests. The 165,698 national requests consisted of 9,234 (5.6%) single requests, 949 (0.6%) exploratory session requests and 155,515 (93.9%) usage session requests.

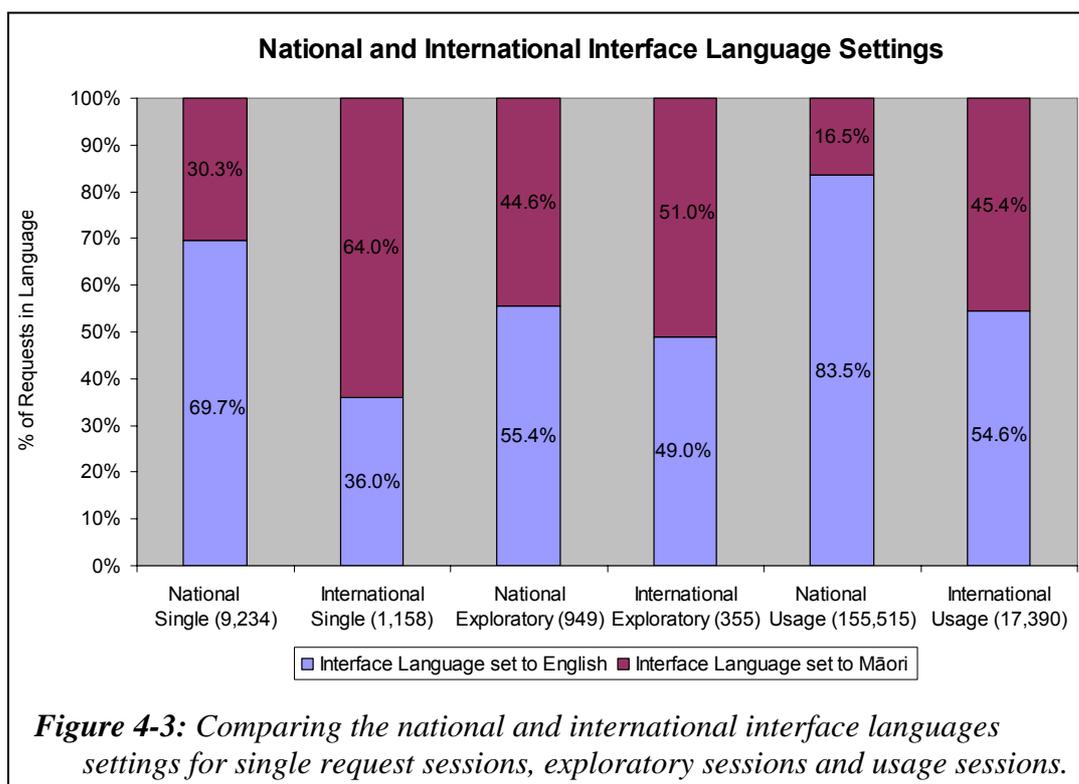
The international requests had a significantly higher proportion of single requests and exploratory session requests. This higher proportion indicates that international clients, having requested a Niupepa interface page, were less likely to then go on and use the Niupepa digital library. As the Niupepa collection presents historic information about Aotearoa it seems reasonable that this information is less interesting to international clients.



4.1.3 Did International Clients Use Te Reo Māori?

Having separated the requests into types of sessions the TLA process was then able to count how many requests were submitted with the Niupepa interface set to te reo Māori and how many requests were submitted with the Niupepa interface set to English. The counts of interface language settings in international sessions and interface language settings in national sessions are displayed in *Figure 4-3*.

Figure 4-3 shows that te reo Māori was used proportionally more often by international clients than it was by national clients. This higher preference for te reo Māori was shown by international clients in all session types.



Usage sessions are the most important to evaluate as they comprise the greatest proportion and give insights into how clients actually used the Niupepa interface. Only 16.5% of national usage requests were undertaken with the interface language set to te reo Māori. However, almost three times that proportion, 45.4%, of international usage requests are requests undertaken with the interface language set to te reo Māori.

International clients did use the Niupepa digital library in te reo Māori and they were almost three times more likely to use it in te reo Māori than their national counterparts.

4.1.4 When Was Te Reo Māori Used?

It is difficult to determine when international requests were submitted from a chronological perspective. Users may log onto the Niupepa web site from any time zone in the world; however the request time and date recorded in the web

logs is the local time in the time zone where the web server is located. The local time of the web server in Aotearoa will not match the local time of international clients. Consequently, international ‘hour of day’ and ‘day of week’ statistics cannot be generated with any degree of certainty. However, the number of requests per month is not affected to the same degree. The months in which international clients submit requests will be recorded correctly in the web log files for at least 97% of the requests but this validity is probably closer to 99%. This accuracy allows a comparison to be undertaken for requests submitted per month by international and national clients.

Percentages of both international and national usage requests in te reo Māori per month were calculated and are displayed in *Figure 4-4*.

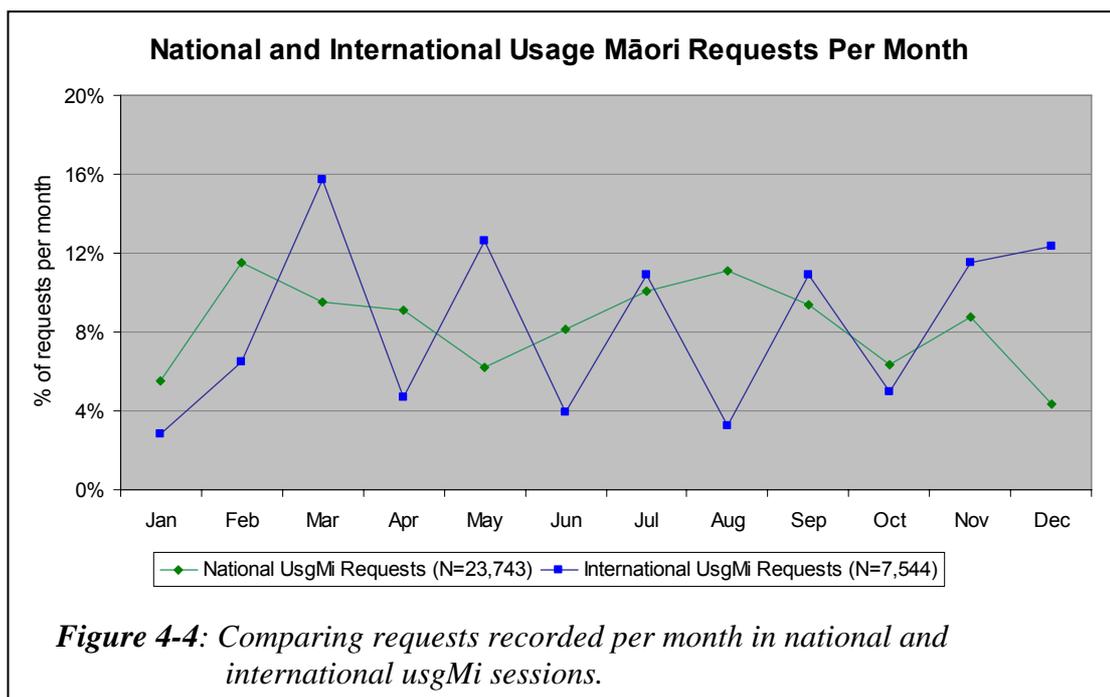


Figure 4-4 clearly shows that the international usgMi requests have monthly cycles of peaks and troughs, whereas the national usgMi requests were more evenly distributed across the year. This raises the question why were the international clients three more likely to undertake requests in te reo Māori in alternative months? The answer can be attributed to the default language setting of the Niupepa web site. In the months of January, March, May, July, September and

November the default language of the Niupepa web site was set to te reo Māori. In the other months the default language was set to English. Aside from the first and last months, the months when the default language was set to te reo Māori are the months during which international clients were three times more likely to use the interface in te reo Māori.

The default language of the web site appears to have had significant impact on which language the clients, in particular international clients, chose to navigate the web site. An extensive examination of this influence is undertaken in Chapter 5.

4.1.5 Summary—Was Te Reo Māori Used?

This section has used statistics generated from TLA to determine if te reo Māori was used by international clients of the Niupepa digital library. First it was established that the Niupepa digital library received 19,303 filtered usage requests from international clients in 2005. International requests comprised approximately 1 in 10 of all usage requests received by the Niupepa digital library. There were a higher proportion of single request sessions and exploratory sessions in international sessions than in national sessions, suggesting that prospective international clients were less likely to use the Niupepa digital library than prospective national clients. However, the international clients were more likely to use the Niupepa digital library in te reo Māori. International clients were almost three times more likely to use the Niupepa digital library in te reo Māori than national clients and were most likely to use the Niupepa interface in te reo Māori when the default language of the interface was set to te reo Māori.

Having established that the Niupepa digital library was used in te reo Māori by international clients, the next stage is to determine how they used te reo Māori and if this usage was different from the usage of te reo Māori by national clients. An analysis of te reo Māori usage by international clients and a comparative analysis of national/international reo Māori usage are detailed in the next four sections.

4.2 International and National Exploratory Behaviours

Clients that requested pages of the Niupepa interface but then did not go on to submit queries or request any documents have been defined as exploratory clients. Requests from exploratory clients comprised less than 1% of the total requests to the Niupepa digital library in 2005. Although this is a low percentage of requests, the data can still be used to compare how national and international clients in te reo Māori explored the Niupepa digital library.

Approximately one quarter of exploratory requests were submitted by international clients and three quarters by national clients. There were 355 requests from 163 exploratory sessions recorded as international requests in 2005. There were 949 requests from 404 exploratory sessions recorded as national requests in 2005. A high proportion of exploratory sessions lasted less than two minutes; 77.5% of national sessions and 84.7% of international sessions. A high proportion of exploratory sessions consisted of just two requests; 79.0% of national sessions and 87.7% of international sessions.

The international and national exploratory sessions were further subdivided by language of use. There were:

- 95 requests in 43 international expMi sessions
- 331 requests in 137 national expMi sessions
- 260 requests in 120 international expEn sessions
- 618 requests in 267 national expEn sessions

Exploratory sessions were categorised as either expMi or expEn. However, they often consisted of just two requests, both for the home page, one with the interface set to te reo Māori and one with the interface set to English. By definition (see §2.4.3), if the exploratory session consisted of just two requests, and the last request was with the interface set to te reo Māori then the session was defined as a usgMi session. Alternatively, if the exploratory session consisted of just two

requests, and the last request was with the interface set to English then the session was defined as a usgEn session. A listing of the international and national exploratory sessions, including which language the interface was set to when requests were submitted, shows:

- 43 international expMi sessions; 79 requests in te reo Māori,
16 requests in English,
- 137 national expMi sessions; 267 requests in te reo Māori,
64 requests in English,
- 120 international expEn sessions; 158 requests in English,
102 requests in te reo Māori,
- 267 national expEn sessions; 462 requests in English,
156 requests in te reo Māori.

The figures listed above indicate that many of the exploratory sessions involved a language switch. TLA was used to examine three further characteristics of exploratory sessions in te reo Māori—the length of expMi sessions, the apparent reading times in expMi sessions and the interface pages requested in expMi sessions. Summaries of these three exploratory characteristics follow.

4.2.1 How Long Were Exploratory Sessions?

Statistics generated by TLA revealed the average lengths of international and national exploratory sessions. The average lengths in time and requests were:

- 3.5 minutes and 2.2 requests in international expMi sessions
- 2.8 minutes and 2.4 requests in national expMi sessions
- 45 seconds and 2.2 requests in international expEn sessions
- 2.5 minutes and 2.3 requests in national expEn sessions

The times listed above indicate that international expMi clients took the longest time, on average 3.5 minutes, to determine that they did not want to use the Niupepa digital library. The international expEn clients were fastest at deciding not to use the Niupepa digital library (average 45 seconds). Just over two requests was the average for all national and international exploratory sessions.

4.2.2 How Long to Read Interface Pages?

The exploratory categories with the longest session lengths were also the categories that exhibited the longest apparent read times of interface pages. The average apparent interface page read times in international and national exploratory sessions were:

- 185.3 seconds in international expMi sessions
- 135.0 seconds in national expMi sessions
- 54.7 seconds in international expEn sessions
- 127.7 seconds in national expEn sessions

The international expMi clients exhibited the longest average read time of interface pages (185.3 seconds). The international expEn clients exhibited the shortest average read time of interface pages (54.7 seconds).

4.2.3 Which Interface Pages Were Requested?

An interface page is a page of the Niupepa web site; it does not refer to a newspaper document in the Niupepa collection. The interface pages requested by international and national expMi clients and international and national expEn clients are listed in *Table 4-1*.

Interface Pages Accessed in Exploratory Sessions								
	expMi				expEn			
	International		National		International		National	
Total pages	95	100.0%	331	100.0%	260	100.0%	618	100.00%
home page	87	91.6%	271	81.9%	243	93.5%	485	78.50%
NZDL home page	0	0.0%	9	2.7%	2	0.8%	58	9.40%
Eng. Abstr. page	7	7.4%	31	9.4%	9	3.5%	26	4.20%
help page	0	0.0%	5	1.5%	2	0.8%	22	3.60%
preferences page	1	1.1%	14	4.2%	4	1.5%	25	4.00%
greenstone page	0	0.0%	0	0.0%	0	0.0%	1	0.20%
unknown page	0	0.0%	1	0.3%	0	0.0%	1	0.20%

Table 4-1: Number and percentages of web site interface pages requested in international and national exploratory sessions.

It can be seen that the majority of the requests, approximately 80% of total national and 92% of total international, were requests for the Niupepa home page. These requests consisted of requests for the home page in te reo Māori, requests for the home page in English and requests for hyperlinks to texts located lower down on the home page. International usgMi clients had a higher preference for requesting the Niupepa home page and, aside from the About English Abstracts page, did not request other Niupepa interface pages.

4.2.4 Summary—Exploratory Sessions

Exploratory sessions in general were short, many simply involving a request for the home page in one language and then a request for the same page in the other language. International expMi sessions were on average slightly longer (3.5 minutes compared with 2.8 minutes) than national expMi sessions and the apparent read times of interface pages were longer (185 seconds compared with 135 seconds). Most of this time was taken reading the Niupepa home page. International usgMi clients did not show the same interest in other pages of the Niupepa interface as national usgMi clients.

4.3 International and National Interface Behaviours

This section analyses how the Niupepa interface was *used* by international reo Māori clients and compares this usage with usage by national reo Māori clients. The purpose of this section is to determine if international clients used the Niupepa interface in te reo Māori differently from national clients.

By definition (see §2.4.2), a usage session is a series of two or more requests from an individual user where each request is no longer than an hour apart and a query is submitted and/or an actual document of the Niupepa Collection is requested. Usage session data is the most useful of all the data analysed as it gives the best insights into how actual users were using the Niupepa interface.

There were 172,905 usage requests recorded in 6,714 sessions for the Niupepa collection in 2005. Of these, 17,390 (10.1%) were recorded as being international usage requests and 155,515 (89.9%) as national usage requests.

The usage sessions requests were categorised into language session types. The counts for the various international and national usage session types were:

- 380 international usgMi sessions consisting of 7,544 requests
- 1,322 national usgMi sessions consisting of 23,743 requests
- 520 international usgEn sessions consisting of 9,340 requests
- 4,432 national usgEn sessions consisting of 128,357 requests
- 14 international usgBi sessions consisting of 506 requests
- 76 national usgBi sessions consisting of 3,415 requests

These figures show that the usage sessions in reo Māori (usgMi sessions) had the highest ratio of international and national requests. Almost a quarter of usgMi requests were from international clients, compared with 1 in 15 of usgEn requests and 1 in 8 of usgBi requests.

Five interface behaviours were examined to compare the international and national usage in te reo Māori of the Niupepa interface. The interface behaviours compared for international and national usgMi sessions were:

- the length of sessions
- the apparent read times of interface pages
- the number of return sessions
- the interface pages requested
- the frequency of interface language switching.

These five interface behaviours are described in the next five sub-sections.

4.3.1 How Long Were Usage Sessions?

Two measurements were used to define length of usage sessions, duration of time (minutes) and number of requests. Statistics generated from TLA were used to compare national and international usage session lengths of the Niupepa digital library in 2005. These comparisons are discussed for both measurements.

How Long in Time Were Usage Sessions?

The durations (in minutes) of the national usgMi sessions were compared with the durations of the international usgMi sessions. The durations of the usage sessions were grouped into time intervals of less than 1 minute, 1–2 minutes, 3–9 minutes, 10–29 minutes, 30–59 minutes and 60+ minutes. The count of sessions that fitted into these categories was collated for national and international sessions. Chi square analysis was undertaken on these counts. The counts and the p values of the chi square analysis are shown in *Table 4-2*.

International and National Duration (in Minutes) of Usage Sessions						
Minutes per session	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Sessions	380	1322	520	4432	14	76
<1	82	374	131	903	0	0
1–2	101	227	146	905	0	4
3–9	89	226	102	798	2	9
10–29	80	234	70	848	4	23
30–59	17	151	47	506	7	25
60+	11	110	24	472	1	15
p values	2.8 * 10 ⁻¹⁰		8.9 * 10 ⁻⁹		Not applicable	

Table 4-2: International and national comparisons for duration groupings of usgMi, usgEn and usgBi sessions showing chi square p values.

A p value of $2.8 * 10^{-10}$ indicates a significant statistical difference in national and international usgMi session durations. Similarly, a p value of $8.9 * 10^{-9}$ indicates a significant statistical difference in national and international usgEn session durations. The low number of usgBi sessions does not give accurate chi square analysis figures so this calculation was not used for usgBi sessions.

Summary statistics were calculated to highlight the international and national duration differences of usgMi, usgEn and usgBi sessions. These summary statistics are displayed in *Table 4-3*.

Table 4-3 indicates that international usgMi sessions on average were completed in almost half the time of national usgMi sessions (10.7 minutes compared with 18.1 minutes). This difference in session duration is statistically significant, as confirmed by the p value displayed in *Table 4-2* ($2.8 * 10^{-10}$). *Table 4-3* also shows that international usgEn sessions were, on average, shorter than national usgEn sessions (13.4 minutes compared with 21.7 minutes) and international usgBi sessions averaged shorter session durations than national usgBi sessions (30.7 compared with 46.0 minutes).

Summary International and National Duration (Minutes) of Usage Session						
Minutes	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Average	10.7	18.1	13.4	21.7	30.7	46.0
Standard Deviation	20.0	32.6	27.2	37.0	17.0	50.3
Most Common	0.0	0.0	0.0	0	11.0	21.0
Least	0	0	0	0	4	3
Lower Quartile	1	0	0	1	16	17
Median	4	5	3	6	33.5	34
Upper Quartile	12	22	13	26	40	52
Most	194	333	239	356	68	256

Table 4-3: International and national comparisons of statistical summaries for durations (in minutes) of usgMi, usgEn and usgBi sessions.

How Long in Number of Requests?

The lengths of usage sessions in requests were also compared between international and national usgMi sessions. The counts of requests per session were grouped into the following intervals 2 requests, 3–4 requests, 5–8 requests, 9–19 requests, 20–39 requests and 40+ requests. Chi square analysis was undertaken on these counts. The counts and the p values of the chi square analysis are displayed in *Table 4-4*.

A p value of 5.1×10^{-4} indicates a significant statistical difference in the count of requests in national and international usgMi sessions. Similarly, a p value of 3.6×10^{-13} indicates a significant statistical difference in the count of requests in national and international usgEn sessions. The low number of usgBi sessions would not give accurate chi square analysis figures so this calculation was not undertaken for usgBi sessions.

International and National Length in Requests of Usage Sessions						
Requests per session	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Sessions	380	1322	520	4432	14	76
2	36	205	51	367	n/a	n/a
3–4	31	185	68	451	n/a	n/a
5–8	90	267	147	780	n/a	n/a
9–19	107	337	140	1179	4	16
20–39	62	185	63	741	5	30
40+	54	143	51	914	5	30
p values	$5.1 * 10^{-4}$		$3.6 * 10^{-13}$		Not applicable	

Table 4-4: International and national comparisons of request groupings in usgMi, usgEn and usgBi sessions with chi square p values.

Summary statistics were calculated to determine exactly what the differences were in international and national usgMi sessions lengths (in request count). These summary statistics are displayed in *Table 4-5*.

International and National Summary of Session Lengths (Requests)						
Requests per session	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Average	19.9	18.0	18.0	29.0	36.1	44.9
Standard Deviation	23.6	27.0	27.5	43.4	21.3	37.8
Most Common	2.0	2	2.0	2.0	10.0	32.0
Least Requests	2	2	2	2	10	13.0
Lower Quartile	6	4	5	6	20	22
Median	11	9	8	13	31	31.5
Upper Quartile	24	20	17	34	53	55
Most requests	192	245	254	598	74	239

Table 4-5: International and national comparisons of statistical summaries for length (in requests) of usgMi, usgEn and usgBi usage sessions.

Table 4-5 indicates that international usgMi sessions averaged more requests (19.9) than national usgMi sessions (18.0). Although this does not appear to be a great difference, the p value displayed in *Table 4-4* ($5.1 * 10^{-4}$) indicates that the difference is statistically significant.

UsgEn sessions and usgBi sessions were different from usgMi sessions in that the average lengths of sessions by request count was longer for national than international clients.

UsgBi sessions were twice as long as the other sessions and had twice as many requests. One of the criteria for defining a usgBi session is that the sessions must consist of at least 10 requests, the result of which means that only long sessions are defined in this category.

4.3.2 How Long to Read Interface Pages?

An apparent read time of interface page is the length of time from when one page of the interface is requested until a subsequent page or document is requested. It is called an apparent read time because we are unable to be certain that the user was *only* engaged with the interface during this time. Statistical summaries of apparent read times in international and national usgMi sessions are displayed in *Table 4-6*. Also listed are apparent read times in international and national usgEn sessions and in international and national usgBi sessions.

It appears that the international usgMi clients took half as long to read a page as the national usgMi clients. A chi square statistical analysis comparing national and international usgMi apparent read times returns a p value of $4.2 * 10^{-32}$. This value implies statistical significant difference.

Apparent Read Times of Interface Pages in Usage Sessions						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Count	387	952	713	5097	93	448
Average (seconds)	27.3	61.4	58.1	47.4	74.6	71.3
Standard Deviation	104.6	179.4	212.4	183.4	228.9	221.8
Mode (seconds)	4	7	4	4	4	4
Shortest (seconds)	0	0	0	0	3	0
Lower Quartile (seconds)	4	6	6	6	7	6
Median (seconds)	5	14	13	12	18	14
Upper Quartile (seconds)	10	40	33	32	39	40
Longest (seconds)	1156	2131	2778	3237	1871	2367
p value	$4.2 * 10^{-32}$		0.61		0.57	

Table 4-6: International and national statistical comparisons of apparent read times of Niupepa interface pages in usgMi, usgEn and usgBi sessions.

While the international and national usgEn apparent read times appear different, with averages of 47.4 and 58.1 seconds, the p value of 0.61 suggests that these are not significantly different. The international and national usgBi sessions, with a p value of 0.57, also have apparent read times that are not significantly different. These p values indicate that if clients were using the interface in English, or reo Māori and English (bilingual use), there was no significant international and national difference in how long the interface pages were being read.

4.3.3 How Many Clients Returned?

Comparisons were made with international return usgMi sessions and national return usgMi sessions to determine if location was a determining factor in whether usgMi clients returned to use the Niupepa digital library. In 2005 there were at least 22 (6.2%) international usgMi clients that undertook more than one session with the Niupepa digital library. There were also at least 60 (4.9%) national usgMi clients that undertook more than one session with the Niupepa digital library. The

international return usgMi clients requested 7,544 (9.7%) pages and documents while national return clients requested 23,743 (15.1%) pages and documents.

Two conclusions can be drawn from these figures. First, the national/international location was not a determining factor in the number of users who returned to use the Niupepa digital library. Second, the national clients who did return to use the Niupepa digital library requested significantly more pages than the international clients who returned. This difference is highlighted when we consider how many requests were submitted per return client. The international usgMi clients requested 33.2 requests per return client. The national usgMi clients requested 59.8 requests per return client. The national usgMi return clients requested almost twice as many pages as the international usgMi return clients.

4.3.4 Which Interface Pages Were Requested?

The percentage of international and national interface pages requested from usgMi, usgEn and usgBi sessions are listed in *Table 4-7*.

When the interface pages requested in the usage sessions (*Table 4-7*) were compared with the interface pages requested in the exploratory sessions (*Table 4-1*) it shows that usage sessions had a higher percentage of requests for the preferences pages. This inclination is understandable as active users were altering the preferences to suit their browsing and searching strategies.

The international usgMi clients were requesting a higher percentage of home pages and preference pages than the national usgMi clients. However, this difference is only 5.1% and 1.3% respectively, a difference that is not statistically significant. The international and national usgEn clients were requesting similar interface pages as the usgMi clients. The usgBi international requests were too few to draw conclusions.

International and National Interface Pages Requested in Usage Sessions						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total pages	407	983	9340	128357	93	458
niupepa home page	84.8%	79.6%	86.7%	80.3%	65.6%	76.9%
NZDL home page	0.0%	1.8%	0.1%	2.2%	0.0%	2.0%
About Eng. Abstract	3.4%	7.0%	4.0%	6.0%	1.1%	5.5%
help page	2.7%	3.8%	2.8%	2.4%	6.5%	4.1%
preferences page	9.1%	7.8%	6.3%	9.0%	26.9%	11.6%
greenstone page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
unknown page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 4-7: International and national comparison of percentages of Niupepa web site interface pages requested in usage sessions.

4.3.5 How Often Was the Language Switched?

Solid conclusions cannot be drawn in comparisons between national and international language switching in usage sessions as there was not enough international language switching data available. Nationally there were 1,512 language switches in 1093 usage sessions; internationally only 206 language switches in 277 usage sessions. Summary international and national switching counts and percentages are displayed in *Table 4-8*.

International and National Switching Summaries						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Number of switches	33	160	187	1057	57	295
reo Māori to English	60.6%	36.9%	90.4%	84.1%	57.9%	55.9%
English to reo Māori	39.4%	63.1%	9.6%	15.9%	42.1%	44.1%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
sessions with switch	6.1%	9.0%	32.5%	20.3%	100.0%	100.0%

Table 4-8: Comparing national and international switching counts and percentages for usgMi, usgEn and usgBi sessions.

It appears that international usgMi sessions had a higher preference for switching from te reo Māori to English (60.6%), whereas national usgMi sessions had a higher preference for switching the interface language from English to te reo Māori (63.1%). International and national usgEn sessions both showed a strong preference to switch from te reo Māori to English (90.4% and 84.1%). International and national usgBi sessions both showed a slight preference to switch from te reo Māori to English (57.9% and 55.9%).

Again, while these assertions can be raised, given the data that has been calculated in *Table 4-8* they cannot be confirmed with any degree of certainty as the count of international switches is too low.

4.3.6 Summary—Interface Behaviours

This section analysed how the Niupepa interface was used by international reo Māori clients and national reo Māori clients. Statistics derived from TLA were used to detail five aspects of how international usgMi clients used the Niupepa interface differently from national usgMi clients.

The statistics revealed that international usgMi sessions averaged 10 minutes, almost half the duration of national usgMi sessions. Although the length in time was less, the usgMi international sessions averaged two more requests per session than the national usgMi sessions. This meant that the international usgMi clients were reading pages more quickly than the national usgMi clients, on average twice as quick. There did not appear to be a significant difference in the number of international and national usgMi clients returning to use the Niupepa collection; however, the national usgMi clients that did return requested almost twice as many pages as the international usgMi clients who undertook return sessions. The international usgMi clients appeared to request a similar proportion of interface pages as the national usgMi clients. While there were too few international usgMi interface language switches (only 33) to draw any firm conclusions, it did appear that the international usgMi sessions had a higher preference for switching the

interface language from te reo Māori to English, whereas national usgMi sessions had a higher preference for switching from English to te reo Māori.

This brief summary completes the comparison of how international and national usgMi clients used the Niupepa interface in 2005. There are two further parts in the comparative analysis of how the Niupepa digital library was used by international and national usgMi clients: a comparative analysis of how documents were accessed and a comparative analysis of how queries were undertaken. The comparative analysis on how documents were accessed by international and national usgMi clients begins in the following section.

4.4 International and National Document Access Behaviours

This section compares how international reo Māori clients and national reo Māori clients accessed documents. A document is defined as a resource within the Niupepa digital library that represents real content, rather than something that provides access to content, such as an interface page. Typically this resource is a newspaper page but it could also be a request for an English Abstract or a request for a Bibliographic Commentary.

Statistics generated from TLA were used to compare four document behaviours of international and national usgMi clients. Document behaviours compared were:

- the methods used to access documents
- the document formats favoured
- the document language favoured
- the apparent read times of documents.

The comparative analysis of these four document behaviours comprises the next four sub-sections.

4.4.1 How Were Documents Accessed?

Three methods were available to clients of the Niupepa digital library accessing documents. The Niupepa documents could be selected from a list of results following a query activity, from a list of publication series and from a list of articles published in a certain date. TLA indicated the methods used to access documents by international and national usgMi clients. The percentages used are displayed *Table 4-9*.

International and National Document Access						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total Documents	5083	17,165	5,955	96,085	224	2,048
From query	84.5%	75.7%	71.5%	82.2%	74.1%	69.8%
From series	10.4%	13.7%	18.0%	5.9%	14.7%	9.8%
From date	2.1%	8.9%	7.1%	9.8%	0.9%	16.8%
Uncertain	3.0%	1.7%	3.5%	2.1%	10.3%	3.7%
p value	2.1*10⁻⁷⁵		7.1*10⁻³⁰⁴		8.1*10⁻¹³	

Table 4-9: *International and national comparison by percentage of how documents were requested by usgMi, usgEn and usgBi clients.*

The query activity was the most popular method of accessing pages, both nationally and internationally for all three usage session types. The very low p values indicate that the differences in how documents were requested by the national and international usage sessions are statistically significant for all types of usage sessions.

The international usgMi clients had a higher preference (84.5%) for using the query facility than the national usgMi clients (75.7%). The national usgMi clients had higher preferences for accessing documents by using the series list and the date facility. The reverse was true for usgEn clients, where national usgEn clients had the highest preference for using a query to access documents.

4.4.2 Which Document Formats Were Favoured?

The newspaper documents stored in the Niupepa collection can be presented to the user in one of three formats, extracted text, full size image, and a preview image. *Table 4-10* lists the numbers and percentages of these three formats with a fourth listing, those requests where it was unclear which page format was requested.

International and National Formats of Requested Documents								
	usgMi		usgEn		usgBi		Total	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total	407	17,165	5,955	96,085	224	2,048	1,262	115,298
Ext. Txt	82.8%	55.5%	85.6%	53.0%	84.4%	51.2%	84.3%	53.4%
Full Size	12.1%	39.5%	8.8%	40.1%	10.3%	38.7%	10.3%	40.0%
Preview	3.5%	4.8%	3.6%	6.4%	3.6%	9.9%	3.6%	6.2%
Undefined	1.6%	0.3%	1.9%	0.5%	1.8%	0.1%	1.8%	0.4%

Table 4-10: International and national comparisons of document types requested by usgMi, usgEn and usgBi clients.

The extracted text format was the most requested format by all client types in both international and national sessions. However, the amount that it was preferred by international and national clients was significantly different. International usgMi clients requested the extracted text format in 82.8% of all document requests, whereas the national usgMi clients requested this format in only 55.5% of documents requests. International usgMi clients exhibited a lesser preference for requesting documents in the full size image format, 12.1% of document requested, as opposed to 39.5% of document requested by national usgMi clients.

The preference differences exhibited by international and national usgMi clients were similar to the preference differences shown by international and national usgEn which, in turn, were similar to those exhibited by international and national usgBi clients.

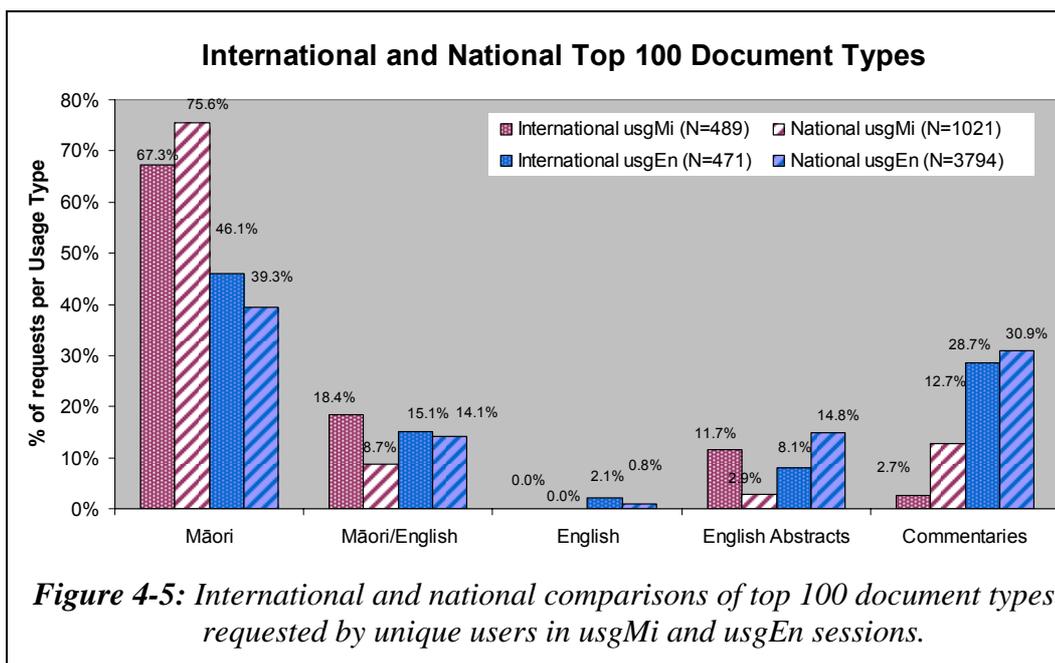
4.4.3 Which Document Types Were Favoured?

Five different types of documents were available to Niupepa clients. These documents consisted of; newspapers written in te reo Māori, newspapers in parallel texts of te reo Māori and English, newspapers written in English, abstracts in English and commentary information written in both English and te reo Māori. A list of document quantity is given in §3.5.3. International and national requests for these document types were compared to determine if international usgMi clients requested different document types from national usgMi clients.

Rather than simply counting the document types requested, the document types requested by unique users were counted. Unique users give a better picture of what a wider range of people accessed and limits the influence in the statistics of the heavy users. It was also decided to collate only the top 100 unique pages requested as limiting the analysis to this number gives a better indication of users' initial preferences, again reducing the influence in the statistics of heavy users of the web site. The data collated in this part of the analysis consisted of:

- 489 unique user requests for 105 pages in international usgMi sessions
- 1,798 unique user requests for 106 pages in national usgMi sessions
- 471 unique user requests for 74 pages in international usgEn sessions
- 3794 unique user requests for 99 pages in national usgEn sessions
- 13 unique user requests for 6 pages in international usgBi sessions
- 233 unique user requests for 91 pages in national usgBi sessions

The usgBi session requests consisted of all the requests for pages by two or more unique users. These counts were considered too low to return valid data so the usgBi sessions were not compared in this particular analysis. The usgMi and usgEn document request counts were calculated as percentages, so that they could be compared, and are graphed in *Figure 4-5*.



Some differences can easily be seen in *Figure 4-5*. International and national usgMi sessions had a stronger preference to request newspaper documents written in te reo Māori. International and national usgEn sessions had a stronger preference to request the commentary information written in English.

The international usgMi clients had stronger preferences than the national usgMi clients to request the parallel reo Māori and English newspaper documents and the English Abstracts. On the other hand, the national usgMi clients had stronger preferences than the international usgMi clients to request the reo Māori newspaper documents and the Bibliographic Commentaries.

The differences noted between the international and national document format preferences of usgMi clients were not apparent to the same extent with international and national usgEn clients.

4.4.4 How Long to Read Documents?

The apparent read times (as explained in §2.5.1) of newspaper documents were compared for international and national usgMi clients. The average apparent read times of full size documents, of extracted text documents and of preview documents were calculated and are displayed for the various usage sessions in *Table 4-11*.

All average reading times in seconds	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
All formats	43.5	79.8	59.2	55.0	57.7	72.5
Full Size Image	77.1	115.1	111.6	93.4	113.0	120.2
Extracted Text	39.3	59.1	55.5	30.3	54.2	46.6
Preview Image	35.9	35.4	37.7	26.3	14.9	24.4

Table 4-11: International and national average apparent read times (in seconds) of documents in usgMi, usgEn and usgBi sessions.

In all types of usage sessions the full size images had the longest average apparent read time. The second longest read averages in all sessions were the extracted text documents. The averages also indicated that the shortest documents to read in all usage sessions were the preview images.

In general the international usgMi clients took less time to read documents than the national usgMi clients. The international usgMi apparent read times were quicker than national usgMi apparent read times for full size images and for extracted text documents and were very similar for preview images.

The reverse was true for usgEn clients. The international usgEn clients averaged longer apparent read times than national usgEn clients for all types of documents.

4.4.5 Summary—Document Access Behaviours

This section compared how international and national reo Māori clients accessed documents. Statistics generated from TLA were used to compare four document access behaviours exhibited by international and national usgMi clients.

The comparisons revealed that using the query facility was the favoured method used to select documents, a method favoured more by international than by national usgMi clients. The extracted text format of documents was preferred by usgMi clients; however, this format was preferred significantly more by international usgMi clients than by national usgMi clients. International usgMi clients displayed higher preferences for requesting parallel reo Māori/English newspaper documents and English Abstracts, whereas national usgMi clients displayed higher preferences for requesting reo Māori newspaper documents and Bibliographic Commentaries. International usgMi clients exhibited quicker apparent read times than national usgMi clients for all types of documents requested.

This brief summary concludes the comparison of how international and national reo Māori clients accessed documents. The comparative analysis of how the Niupepa digital library was used by international and national usgMi clients consists of three parts, a comparative analysis on how the interface was used, a comparative analysis of how documents were accessed and a comparative analysis of how queries were undertaken. The third and final part, how queries were undertaken by international and national usgMi clients, begins in the following section.

4.5 International and National Query Behaviours

This section analyses and compares query behaviours of international and national usage sessions in te reo Māori. A query is defined as the activity undertaken by a client that involves using the search facility of the Niupepa digital library to find a term or string of terms in documents stored in the Niupepa digital library.

Perl scripts have been used in the TLA process to generate query behaviours from the NZDL National and NZDL International web log datasets. The comparative analysis was undertaken on nine query behaviours:

- the number of queries submitted
- the number of query terms used in queries
- the topics queried
- the language of the query terms
- the use of query parameters
- the use of query operators
- the number of documents requested following a query
- the format of documents requested following a query
- the apparent document read times following a query.

These nine query behaviours comprise the next nine sub-sections and detail a comprehensive comparative analysis of how international and national reo Māori clients queried the Niupepa digital library.

4.5.1 How Many Queries Were Submitted?

The numbers of queries submitted by international usgMi clients were compared with the numbers of queries submitted by national usgMi clients. These counts were calculated as percentages and are displayed in *Table 4-12*. Also displayed in *Table 4-12* are the percentages of queries submitted by international and national usgEn clients and by international and national usgBi clients.

International and National Sessions That Contain Queries						
	usgMi		usgEn		usgBi	
Queries/Session	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total Sessions	380	1322	520	4432	14	76
no queries	30.0%	46.4%	46.7%	31.4%	21.4%	19.7%
1–2 queries	35.3%	27.4%	28.8%	32.9%	14.3%	11.8%
3+ queries	34.7%	26.2%	24.4%	35.7%	64.3%	68.4%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 4-12: International and national comparisons of number of sessions that contain queries in usgMi, usgEn and usgBi sessions.

International usgMi clients were more likely to submit queries than national usgMi clients. International usgMi clients submitted queries in 70.0% of their sessions, while national usgMi clients submitted queries in only 53.6% of their sessions. When usgMi sessions did involve a query activity, half of these sessions comprised just one or two queries and the remainder comprised three or more queries.

The likelihood of an international and national client submitting a query was reversed when the interface language was set to English. International usgEn clients submitted queries in only 53.2% of their sessions, while national usgEn clients submitted queries in 68.6% of their sessions. The international usgBi sessions appeared to submit queries in a similar proportion to the national usgBi sessions; however, because of the low international usgBi session count (there were only 14) this conclusion remains indeterminate.

4.5.2 How Many Query Terms Were Submitted?

A count was taken of the number of query terms submitted with each international and national usgMi query. The percentages, the number of queries, and the chi square p values are displayed in *Table 4-13*.

International and National Number of Query Terms per Query						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total Queries	1399	3368	1324	15,907	82	422
1 term/query	42.5%	48.6%	58.3%	48.9%	50.0%	46.4%
2 terms/query	30.3%	29.3%	24.8%	32.4%	36.6%	32.2%
3+ terms/query	27.2%	22.0%	16.8%	18.7%	13.4%	21.3%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
p value	$6.7 * 10^{-5}$		$1.2 * 10^{-10}$		0.26	

Table 4-13: International and national percentage comparisons for the number of terms in a query for usgMi, usgEn and usgBi sessions.

A comparison of the national and international percentages in the usgMi sessions seems to suggest that the proportions are similar. However, with a chi square p value of $6.7 * 10^{-5}$ there is a significant difference in the number of query terms submitted in international and national usgMi queries. International usgMi queries contained a higher proportion of queries with three or more query terms. National usgMi queries contain a higher ratio of queries with just one query term.

The reverse was true for usgEn sessions. International usgEn queries contained a higher proportion of queries with just one query term, while national usgEn queries contained a higher proportion of queries with three or more query terms.

4.5.3 What Topics Were Queried?

The topic categories of the query terms were analysed to see if there was a difference between query terms submitted internationally and those submitted nationally. The topic categories were not analysed to the language of session level as the counts of terms became too low to draw any meaningful conclusions. Thus, international and national comparisons were made but international usgMi/national usgMi comparisons were not analysed.

International and National Query terms		
	International Terms	National Terms
1	haka (28)	whakatauki (68)
2	harakeke (21)	makaka (53)
3	waiata (15)	Whakaki (48)
4	koingo (12)	awaitaia (41)
5	tutakangahau (12)	matariki (36)
6	aroha (10)	tioriori (35)
7	mangamaunu (10)	waiata (35)
8	hawaii (9)	parihaka (31)
9	hetet (9)	Parihaka (28)
10	churton (8)	Te wharepu (28)
11	Hetita (8)	treaty (27)
12	R.T.K. (8)	oruanui (26)
13	India (8)	Tapiri (24)
14	Immigration Lists (7)	motakotako (24)
15	pounamu (7)	tawhiao (24)
16	urewera (7)	treaty of waitangi (23)
17	davis (7)	Ratana (23)
18	Immigration Ships from Britain (7)	kingitanga (23)
19	tuhoe (7)	arawa (22)
20	Newspapers (7)	"friend (22) ⁱ
21	india (7)	utiku potaka (22)
22	whiro tirea (6)	pipiwharauoa (20)
23	dog tax (6)	rangatiratanga (19)
24	whatu (6)	Tuwhenua (19)
25	tauanui (6)	Chinese (19)
26	taniwha (6)	ngapuhi whakatauki (19)
27	"tukua te rangi" (6)	

Table 4-14: *The query terms (with counts) most commonly submitted by international and national clients.*

ⁱ This search term included an opening speech mark (but not a closing one).

Internationally, 2,803 queries were submitted which consisted of 2,038 different query strings. Nationally, 19,668 queries were submitted which consisted of 12,204 different query strings. Query terms that fell into the top 25 most often searched terms are displayed in *Table 4-14*, with the number of times that the term was requested written in brackets after the term.

Of the query terms listed above, only one term, waiata (*song*), is listed in both the national and international top 25 query terms.

As mentioned in §3.6.3, the topics of the search terms were categorised so that they could be compared to determine international and national differences. The top 100 unique query terms were examined and categorised into the topic headings. For international usage sessions this involved 132 terms that were queried 614 times. For national usage sessions this involved 103 terms that were queried 1761 times. These terms were categorised into the topic categories and then the percentages of terms per category were calculated. These query topic percentages are displayed in *Figure 4-6*.

Figure 4-6 shows that the topic categories searched for by international and national clients were similar. The topic most searched for was names of people followed by names of places. The national clients showed significantly higher preferences for searching about tribes and nationalities, and publications.

In the top 100 unique query terms, only five terms were common to searches undertaken by both international and national clients. Those common query terms were, in order of frequency, waiata [*song*] (15 international requests, 35 national requests), haka [*dance with postures*] (28 international requests, 15 national requests), tuhoe [*a tribe located in the Central North Island of Aotearoa*] (15 international requests, 7 national requests), māori (4 international requests, 12 national requests), and kingitanga [*the kingship*] (3 international requests, 10 national requests).

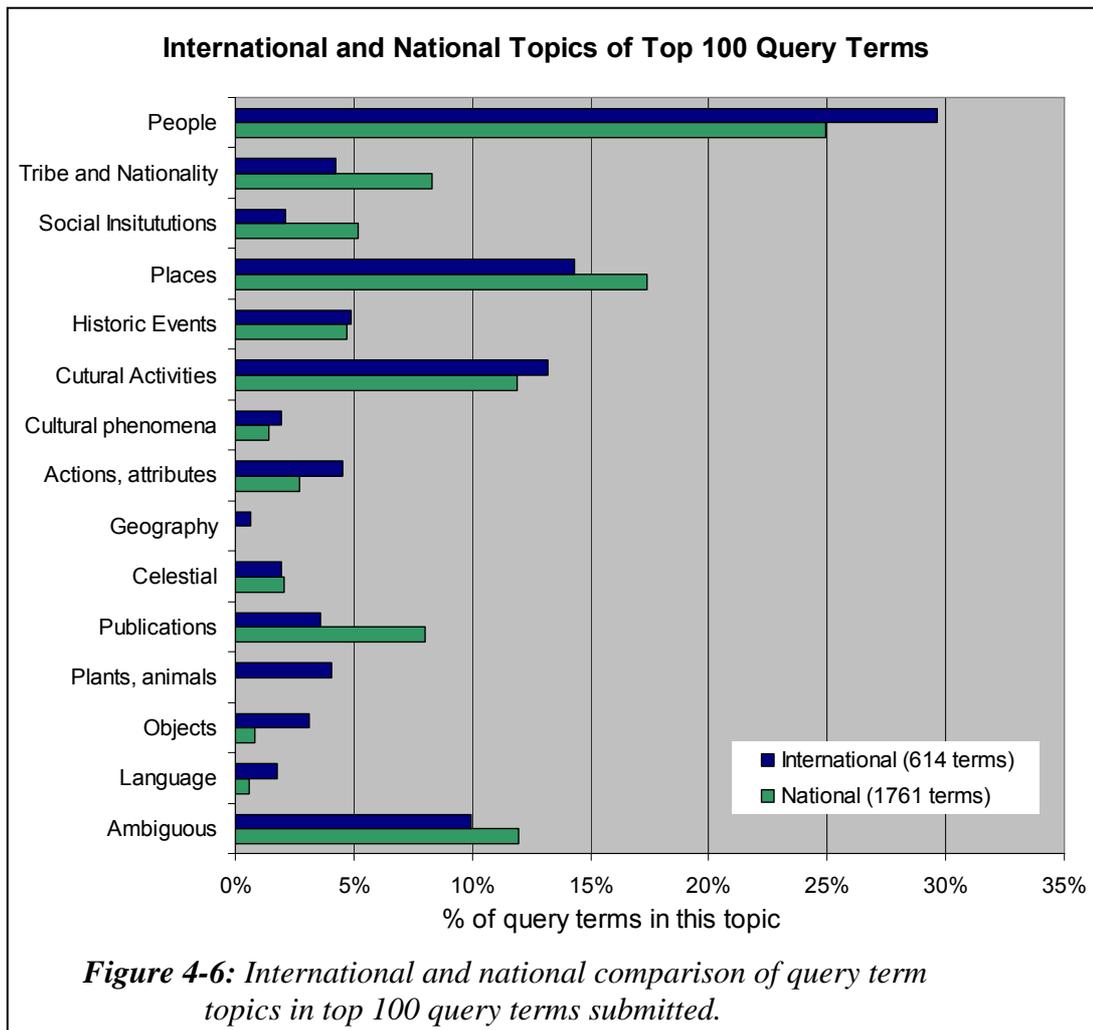


Figure 4-6 suggests the categories searched by international and national users are similar. A comparison of the top 100 query terms shows that there was a 95% difference in the terms used by international and national clients. Different query terms can be expected in the people and places categories as clients will search for names and places that are individual to them. However, similar query terms could be expected for the other categories. The fact that these query terms are not the same suggests that international and national clients were using different vocabularies to retrieve similar topics of information.

4.5.4 What Was the Language Of the Query Terms?

The international and national datasets were examined to determine what language was used in the query terms submitted to the Niupepa digital library in 2005. In total there were 14,160 disparate query terms submitted in 22,471 queries. Most query terms were written in te reo Māori, with almost all of the remaining terms written in English. Seven query terms consisted of numbers, presumably dates and, for the purpose of this analysis, these terms were not considered as reo Māori or English terms.

The language of the most common query terms submitted by international usgMi clients was compared with the language of the most common query terms submitted by national usgMi clients. There were 167 international usgMi query terms examined, submitted in 437 queries, with 94 national usgMi query terms examined, submitted in 532 queries.

The international usgMi clients submitted 95.9% of their query terms in te reo Māori. The national usgMi clients submitted 90.0% of their query terms in te reo Māori. These figures show that the international usgMi clients had a higher (5.9%) preference for submitting query terms in te reo Māori.

4.5.5 Were Query Parameters Altered?

Clients who undertook queries had the facility to alter some query parameters on the Preferences page and on the Query page. On the Preference page, clients were able to alter parameters such as query mode, query style, case matching, word endings and search history (see *Figure 3-2*). On the Query page, clients were able to alter whether the results of the search must contain *some* or *all* of the query (see *Figure 3-3*).

The query parameters located on the Preference page were not altered by international or national usgMi clients. Only 0.9% of international usgMi clients and 0.2% of national usgMi clients altered one of these parameters.

The Some/All query parameter located on the Query page, a page used by clients to submit queries, was altered by international and national usgMi users. There were 11.0% of international usgMi clients and 7.6% of national usgMi clients who altered the Some/All query parameter. These percentages indicate that international usgMi clients had a higher preference to alter this query parameter than the national usgMi clients.

In contrast, the reverse preference was noted in international and national usgEn clients. There were 5.9% of international usgEn clients and 10.9% of national usgEn clients who altered the Some/All query parameter. These percentages indicate that national usgEn clients had a higher preference to alter this query parameter than the international usgEn clients.

4.5.6 Were Query Operators Used?

Two types of query operators were available to clients of the Niupepa digital library in 2005; logical query operators (and, or, not, *, +, -) and phrase defining operators (quotation marks). Less than 2% of international and national queries contained the use of logical operators; consequently, the use of these operators was not investigated further. However, it was a different matter for phrase defining operators.

There were 344 queries (out of a total 1,399 queries) submitted by international usgMi clients that included the use of quotation marks to specify phrase searching. This figure represents one in four (24.6%) of all the queries submitted by international usgMi clients and is a considerable proportion. There were 481 queries (out of a total 3,368 queries) submitted by national usgMi clients that included the use of quotation marks. This figure represents one in seven (14.3%) of all national usgMi queries.

Queries submitted by international usgEn clients also showed a higher presence of phrase defining operators (8.9%) than queries submitted by national usgEn clients

(6.2%). The difference observed in international and national usgEn clients was not as large as the difference observed in international and national usgMi clients.

4.5.7 Documents Requested Following a Query

The numbers of documents requested following the submission of a query were examined for international and national clients who used the interface in te reo Māori. There were 1,399 queries submitted by international usgMi clients that led to the request of 4,213 documents; thus, 3.0 documents were requested for every international usgMi query. There were 3,368 queries submitted by national usgMi clients that led to the request of 11,868 documents; thus, 3.5 documents were requested for every national usgMi query. The national usgMi clients requested more documents following a query than the international usgEn clients.

As a comparison, there were 3.2 documents requested following an international usgEn query and 4.7 documents requested following a national usgEn query. These figures indicate that the national usgEn clients requested significantly more documents following a query than the international usgEn clients.

4.5.8 Formats Favoured Following a Query

There were 8,572 international and 88,290 national requests for documents following a query. The percentages of document formats requested by international and national usgMi clients following a query are displayed in *Table 4-15*. Also listed are the percentages of document formats requested by international and national usgEn and usgBi clients.

International and National Document Formats Requested After a Query						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total	4,213	11,868	4193	74,999	166	1,423
Extracted Text	89.0%	62.9%	90.2%	55.0%	90.4%	51.7%
Full Size Image	8.8%	33.0%	6.8%	40.1%	6.6%	36.5%
Preview Image	1.9%	4.0%	2.2%	4.7%	2.4%	11.6%

Table 4-15: International and national percentages of document formats requested following a query.

International usgMi clients requested a higher proportion (89.0%) of extracted text documents following a query than national usgMi clients (62.9%). Alternatively, national usgMi clients requested a higher proportion of full size images (33.0%) and preview images (4.0%) than international usgMi clients (8.8% and 1.9% respectively).

The preference in the international usage sessions to request the higher proportion of extracted text documents and the national usage sessions to request the higher proportion of images following a query was also evident in the usgEn and usgBi sessions.

4.5.9 Document Read Times Following a Query

The apparent read times of documents requested following a query were examined for national and international usgMi sessions. This comparison is similar to the comparison undertaken in §4.4.4; however, in this instance only the documents requested following a query were used to calculate the apparent read times. The average apparent read times for each type of document requested are displayed in *Table 4-16*.

International and National Apparent Read Times after a Query

Average Apparent Read (Seconds)	International usgMi		National usgMi	
	Non Query	After Query	Non Query	After Query
All Documents	58.5	40.5	92.1	74.7
Full Size Image	78.7	76.2	130.6	104.2
Extracted Text	58.1	37.1	49.9	61.5
Preview Image	32.6	43.9	24.4	43.6

Table 4-16: Apparent read times (seconds) by international and national usgMi clients following the submission of a query.

Table 4-16 shows that in general usgMi clients, both nationally and internationally, took less time to read documents following the submission of a query. The decrease in apparent read times of all documents by international and national usgMi clients was approximately 18 seconds.

While both international and national clients displayed a similar decrease in apparent read times for all documents, when the document types were analysed individually there were distinct differences. International usgMi clients took about the same time to read full size images, less time to read extracted text documents and more time to read preview images following a query. National usgMi clients took less time to read full size images, more time to read extracted text documents and more time to read preview images following a query.

4.5.10 Summary—Query Behaviours

This section analysed and compared query behaviours of international and national usgMi clients. Statistics produced by the TLA process were used to examine nine query behaviours exhibited by international and national usgMi clients.

These statistics revealed that international usgMi clients were more likely to submit queries than national usgMi clients. When queries were submitted,

international usgMi clients were more likely to search for multiple query terms. International and national clients searched for similar topics but they also used a different vocabulary of query terms. The international usgMi clients entered a higher percentage of query terms in te reo Māori, were more likely to alter the Some/All query parameter and used the phrase defining operator (the quotation marks) almost twice as often as the national usgMi clients. Both of these behaviours are perhaps a consequence of usgMi clients submitting a higher number of terms per query. Fewer documents were requested by international usgMi clients following a query but they requested a higher proportion of extracted text documents. Generally, both client types spent less time reading documents following a query, except in the case of preview images which showed longer reading times following a query.

A comparative analysis of how international and national usgMi clients *used* the Niupepa digital library has been discussed by examining three behaviours, including query behaviours. Before this chapter is complete, one further comparison merits discussion; how international and national usage differences in te reo Māori compared with international and national usage differences in English. This question is addressed in the following section.

4.6 Te Reo Māori Usage/English Language Usage

The purpose of this chapter has been to determine if, and subsequently define how, te reo Māori was used differently by international and national clients of the Niupepa digital library. Given that international and national usgMi differences were noted, another question springs to mind. Were similar international and national differences displayed by clients who used the Niupepa digital library in English? If the international and national differences occurred in a similar manner in English perhaps these differences are likely to occur for all languages? If this is the case then strategies undertaken to increase the international participation in one language can be applied to other languages.

The comparison of international and national usgMi and usgEn behaviours has been condensed into three sections, a comparison of interface behaviours, a comparison of document access behaviours and a comparison of query behaviours.

4.6.1 Interface Behaviour Differences

International usgEn sessions were shorter than national usgEn sessions in both time and number of requests. The apparent read times of interface pages by international and national usgEn clients were similar. National usgEn clients showed interest in the NZDL web site home page and the Niupepa preferences page whereas the international usgEn clients did not show the same interest in these pages.

When international and national usgEn clients switched the language of the interface, both had a high preference to switch the language to English.

The international and national usgEn interface behaviour comparisons were different from the international and national usgMi interface behaviour comparisons in almost all aspects. The only similar aspect was that international usgMi sessions were shorter in time than national usgMi sessions; all other comparisons showed dissimilar results. The international and national usgMi read times were different, the international and national usgMi clients requested proportionally the same interface pages, and when the international and national usgMi clients switched languages they did so in different ways.

4.6.2 Document Access Behaviour Differences

International usgEn clients had a lower preference than the national usgEn clients for using the query facility to access documents and a subsequently higher preference for selecting documents by browsing the publication series list.

Extracted text documents were the formats preferred more by international usgEn clients than by national usgEn clients, with national usgEn clients displaying a higher preference for viewing full size images. There was little difference in the types of documents preferred by international and national usgEn clients. The international usgEn clients averaged longer apparent read times for all types of documents.

The international and national usgEn document access behaviours were different from the international and national usgMi document access behaviours in three out of four aspects. Preferences for the type of document were the only similarity.

4.6.3 Query Behaviour Differences

International usgEn clients were less likely to submit queries than national usgEn clients. International usgEn queries contained a higher proportion of queries with just one query term, while national usgEn queries contained a higher proportion of queries with two or more query terms. There were 5.9% of international usgEn clients who altered the Some/All query parameter compared with 10.9% of national usgEn clients. Queries submitted by international usgEn clients showed a higher presence of phrase defining operators (8.9%) than queries submitted by national usgEn clients (6.2%). International usgEn clients requested significantly fewer documents following a request than the national usgEn clients (3.2 compared to 4.7). International usgEn clients requested a higher proportion of extracted text documents following a query and a lower proportion of full size and preview.

The international and national usgEn query comparisons were different from the international and national usgMi query comparisons in almost all respects. Of the six international and national usgEn query comparisons discussed, only one, the types of documents requested following a query, was similar to the query behaviours noted in the international and national usgMi comparative analysis.

4.6.4 Summary—Behaviour Differences

The results of analysing the international and national usgEn behaviours show there were distinct differences in behaviours exhibited by international and national usgEn clients. These differences occurred in interface behaviours, document access behaviours and query behaviours. A conclusion that can be drawn is that the information retrieval strategies used by international clients who used the Niupepa digital library in English were different from those used by national clients who used the Niupepa digital library in English.

The differences displayed by international and national usgEn clients were not the same as those displayed by international and national usgMi clients. Thirteen of sixteen international and national usgEn behaviours were either the complete opposite of or significantly different from the international and national behaviours exhibited by usgMi clients. These results suggest that usage of the Niupepa digital library by international clients in English is quite different from usage by international clients in te reo Māori. Consequently, strategies undertaken to increase the international participation in one language cannot be applied to another language.

4.7 Summary

This chapter began by asking if te reo Māori was used by international clients of the Niupepa collection. §4.1 showed that international clients were responsible for only 10% of the requests and were more likely to explore but then not go on to use the Niupepa digital library. However, when international clients did use the Niupepa digital library they were three times more likely to use the digital library in te reo Māori than were national clients.

Exploratory activities of international expMi clients were compared with exploratory activities of national expMi clients in §4.2. International expMi sessions were shown to be 20% longer (in time) with 27% longer apparent read times than the national expMi sessions. International expMi clients did not show the same interest in other pages of the Niupepa interface as did their national counterparts.

Usage activities of international usgMi clients were compared with usage activities of national usgMi clients in three areas; interface behaviours (§4.3), document access behaviours (§4.4) and query behaviours (§4.5). International usgMi sessions tended to be quicker, with more requests but comparatively fewer interface language switches than national usgMi sessions. International usgMi clients were more likely to use queries but requested fewer documents following a query. International usgMi clients requested more extracted text formats and parallel (reo Māori/English) and English language documents than national usgMi clients. The higher proportion of queries undertaken by international usgMi clients included a higher proportion of multiple query terms, more use of the Some/All parameter and the phrase defining operator. These differences were all statistically significant and revealed that te reo Māori was not used in the same way by international clients of the Niupepa digital library as by national clients.

A further question was considered in the preceding section (§4.6): given that te reo Māori was used in different ways by international and national clients, was the English language also used differently? Results from the international and

national usgEn comparison were completely opposite to or significantly different from results from the international and national usgMi comparison in 13 of the 16 behavioural differences analysed. Consequently, international and national differences that occurred in reo Māori usage were not the same as international and national differences in English language usage.

4.7.1 Discussion

The Niupepa web site is likely to hold more interest for national clients, users based in Aotearoa, because of the nature of the content. Historical newspaper documents are more likely to be of interest to people from the country in which the newspapers were published. Further, if the newspapers are mostly written in an indigenous language, then they are of more interest to the country where that indigenous language is spoken. This higher interest has been evident in the statistics generated when comparing national and international requests. Almost 85% of all filtered requests and 90% of all usage requests to the Niupepa web site were requests from (national) clients in Aotearoa.

It was difficult to determine exactly where the international usgMi clients were from. Statistics discussed in §3.4.4 indicated that most international requests submitted by the top 100 IP Addresses were coming from Australia and the United States of America. However, only 2.7% of requests from Australia and 5.9% of requests from USA were catalogued as usgMi requests. A large number of international usgMi requests did appear to originate from the islands of Hawai'i but as only the top 100 IP Addresses were identified it is difficult to confirm if this number was significant.

The international clients who used the Niupepa digital library in te reo Māori have been shown to be quite a different user group than the national clients who used the Niupepa digital library in te reo Māori. The increase in international usgMi sessions when the default language was set to te reo Māori (see *Figure 4-3*) and the relatively low number of interface switches indicated that international usgMi

clients were bilingual users who were appeared confident in using the interface in te reo Māori. The international usgMi clients appeared to be experienced in information retrieval; their preference for queries and their use of query parameters, query operators and multiple query terms indicated they were familiar with query activities. The fewer documents requested following a query suggested that their query activity was more focused and/or more accurate. Their short usage session times but high usage session requests indicated they were precise in obtaining the information that they were seeking. The higher proportion of requests for parallel (reo Māori/English) and English language documents suggested that international usgMi clients were seeking information for English language purposes more often than were national usgMi clients.

The international usgMi clients of the Niupepa digital library appeared to be experienced information retrieval clients who were confident to use te reo Māori. These clients were able to take advantage of electronic reo Māori resources made available on the Internet. Use of te reo Māori by international clients in the Niupepa digital library is an example of how a resource such as the Niupepa digital library can be used to assist in the international usage of an indigenous language.

5 The Effect of the Default Language Setting On Te Reo Māori Usage

This chapter examines how the default language setting of the Niupepa digital library affected use of te reo Māori by Niupepa clients. Did setting the default language to reo Māori (defMi) increase use of the digital library in te reo Māori, as proposed in preliminary research (Keegan & Cunningham, 2005)? Did setting the default language to reo Māori change the usage behaviours of te reo Māori clients who used the Niupepa digital library? In this chapter statistics derived from TLA will be used to seek answers to these questions.

The default language is defined as the language that the Niupepa web site was displayed in when the home page (www.nzdl.org/niupepa) was requested. Default language analysis is possible because throughout 2005 the default language setting of the Niupepa web site was alternated, on a monthly basis, between te reo Māori and English, as described in §2.2.3.

Statistics generated from Transaction Log Analysis (TLA) are used to compare te reo Māori usage for the default language settings of reo Māori and English. These statistics show the information retrieval behaviours of te reo Māori clients were significantly different for the different default language settings. These differences are discussed in sections about the amount of reo Māori use, exploratory behaviours, usage of the Niupepa interface, accessing documents and undertaking queries.

5.1 Did Te Reo Māori Usage Increase?

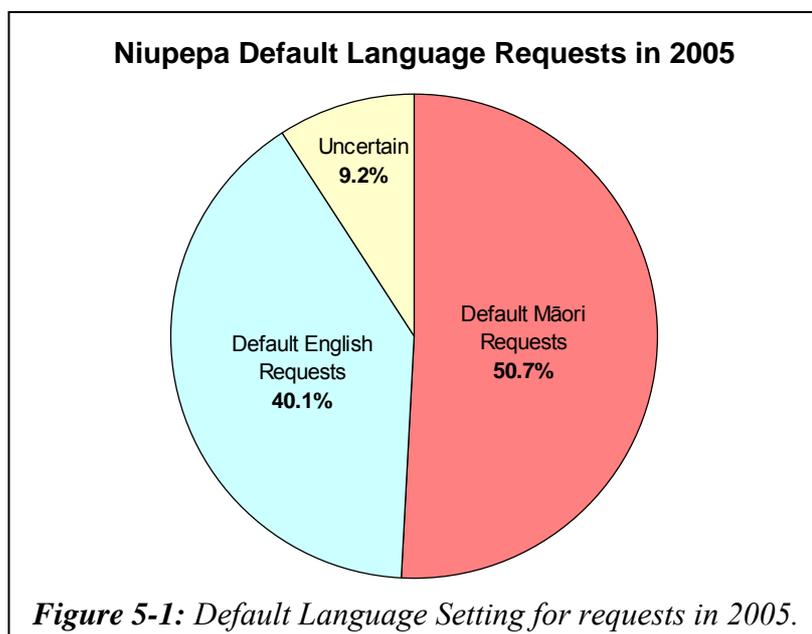
When the default language of a bilingual interface is set to an indigenous language, it can be expected that use of the interface in the indigenous language will increase. Bilingual clients are likely not to change the language of the interface and monolingual clients who do not understand the indigenous language will register at least one request in the indigenous language before the interface is changed to the familiar (non-indigenous) language. But exactly how many more indigenous language requests were there when the default language was set to the indigenous language? Did this setting discourage other (non-indigenous) clients from using the digital library? How does this setting affect the chronological behaviours of clients accessing the digital library? These questions will be addressed in the next five sub-sections.

5.1.1 Effect on the Number of Niupepa Requests

The Niupepa web site received 185,001 filtered requests in 2005. Of these:

- 93,867 were received when reo Māori was the default language (defMi)
- 74,186 were received when English was the default language (defEn)
- 16,948 were received when it was uncertain which interface language was set as the default.

The numbers of requests recorded with the different default language settings have been calculated as percentages and are displayed in *Figure 5-1*.



Interestingly, there were significantly fewer web robot requests when the default interface language was set to te reo Māori. There were 28.8 web robot requests per day with a defMi setting and 75.8 per day with a defEn setting. This difference reveals that web robots submit almost three times more requests to a reo Māori/English bilingual web site when the default language of the web site is set to English.

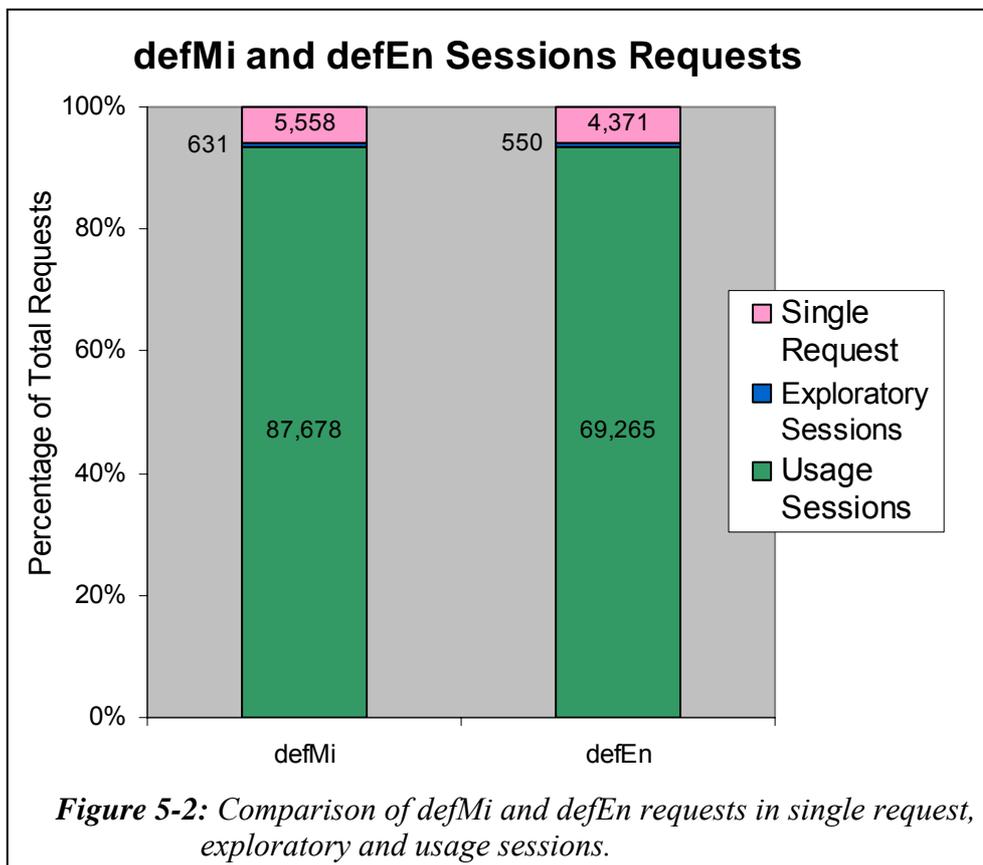
The 93,867 defMi requests were recorded in 198 days at a rate of 474.1 requests per day. The 74,186 defEn requests were recorded in 137 days at a rate 541.5 requests per day. Setting the Niupepa digital library default language to te reo Māori resulted in 12.5% fewer requests overall.

5.1.2 Were Potential Niupepa Clients Discouraged?

Requests recorded when the default language was set to te reo Māori (defMi) and when the default language was set to English (defEn) were classified into single request sessions, exploratory sessions and usage sessions. If the default language

setting significantly affected the number of potential clients then this would be apparent in the proportion of single request, exploratory and usage sessions. A higher proportion of defMi single request and exploratory sessions coupled with a lower proportion of defMi usage sessions would indicate fewer potential clients were using the Niupepa collection because of its default language being set to te reo Māori.

The numbers of requests in defMi and in defEn sessions are displayed together in *Figure 5-2*.



The defMi and defEn showed a remarkable similarity in the proportion of requests per sessions type. There were 5,558 defMi requests (5.92%) and 4371 defEn requests (5.89%) classified as single requests. There were 631 defMi requests (0.67%) classified into 277 defMi exploratory sessions and 550 defEn requests (0.74%) classified into 237 defEn exploratory sessions. There were 87,678 defMi

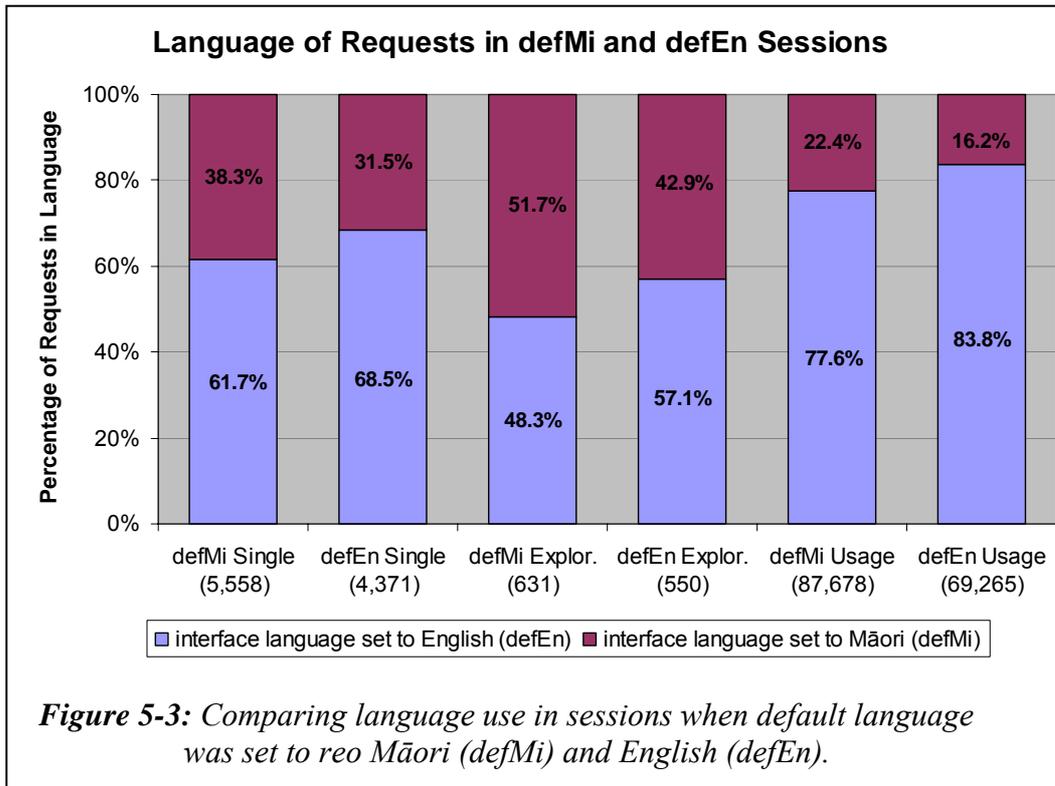
requests (93.41%) classified into 3,238 defMi usage sessions and 69,265 defEn requests (93.37%) classified into 2,971 defEn usage sessions. Chi square analysis of request counts gives a p value of 0.24, indicating no significant difference in the proportions of defMi and defEn session counts.

Setting the default language of the Niupepa digital library to reo Māori did not appear to discourage potential Niupepa clients. There were the same proportion of exploratory/usage sessions by Niupepa clients when the default language was set to reo Māori as there were when the default language was set to English.

5.1.3 Effect on Number of Te Reo Māori Requests

The number of requests submitted in te reo Māori when the default language was set to reo Māori was compared with the number of requests submitted in te reo Māori when the default language was set to English. These numbers have been calculated as percentages and are displayed in *Figure 5-3*.

For the three session types, the proportion of requests in te reo Māori was higher when the default language was set to reo Māori. For single request sessions there was a 6.8% higher proportion of requests in te reo Māori with a defMi setting (38.3% compared with 31.5%). For exploratory sessions there was a 8.8% higher proportion of requests in te reo Māori with a defMi setting (51.7% compared with 42.9%). For usage sessions there were a 6.2% higher proportion of requests in te reo Māori with a defMi setting (22.4% compared with 16.2%). Overall the percentage of requests in te reo Māori increased by 6.2% when the default language of the interface was set to te reo Māori.



As well as the percentage of requests in te reo Māori, the number of requests in te reo Māori was also examined. The requests show an average of 99.1 reo Māori usage requests per day with a defMi setting and 81.9 reo Māori usage requests per day with a defEn setting. There was also an average of 423.7 English usage requests per day with a defEn setting and 343.7 English requests per day with a defMi setting.

Setting the Niupepa digital library default language to te reo Māori resulted in a 21% increase in Niupepa requests in te reo Māori.

5.1.4 Effect on Number of Te Reo Māori Clients

The statistics produced by the TLA process showed counts of how many unique identifiers had been generated for the different session types (§2.4.1 has a discussion on unique identifiers). Unique identifier numbers cannot give an accurate figure of unique client numbers because of behaviours such as disabling

cookies and multiple users sharing single accounts. However, unique identifiers indicate client numbers, which can then be used for comparison purposes.

In the 137 days that the default language of the interface was set to English, 580 unique identifiers were recorded in reo Māori usage sessions. In the 198 days that the default language of the interface was set to reo Māori, 883 unique identifiers were recorded in reo Māori usage sessions. This equates to 4.2 unique identifiers per day with a defEn setting and 4.5 unique identifiers per day with a defMi setting. Setting the default language of the interface to reo Māori did appear to increase the numbers of unique clients in te reo Māori but not to a significant extent.

In contrast, the default language setting did appear to affect the number of usage clients in English by a significant amount. There were 15.8 unique identifiers per day in English usage sessions with a defEn setting but 10.1 unique identifiers per day in English usage sessions with a defMi setting. Setting the default language of the interface to reo Māori appeared to decrease by 36% the numbers of unique clients who used the Niupepa digital library in English.

5.1.5 Effect on When Usage Occurred?

Requests recorded in defMi and defEn usage sessions have been plotted chronologically and are shown in *Figures 5-4, 5-5 and 5-6*. The sessions have been plotted as percentages of requests per session to take into account the variation in frequency of requests between the two default language settings.

Figure 5-4 shows the requests as they occurred in a 24 hour time period. The requests occurred in similar time periods no matter which default language was set. Most of the requests were received during regular 9:00am to 5:00pm work hours, with a small peak again in the evening, probably after dinner time.

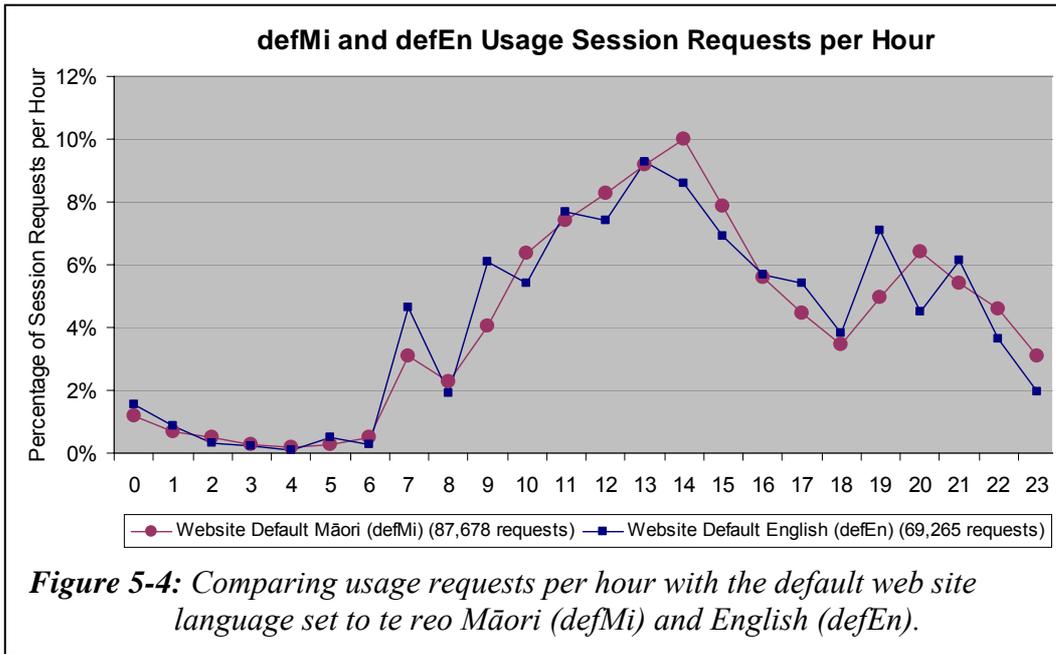
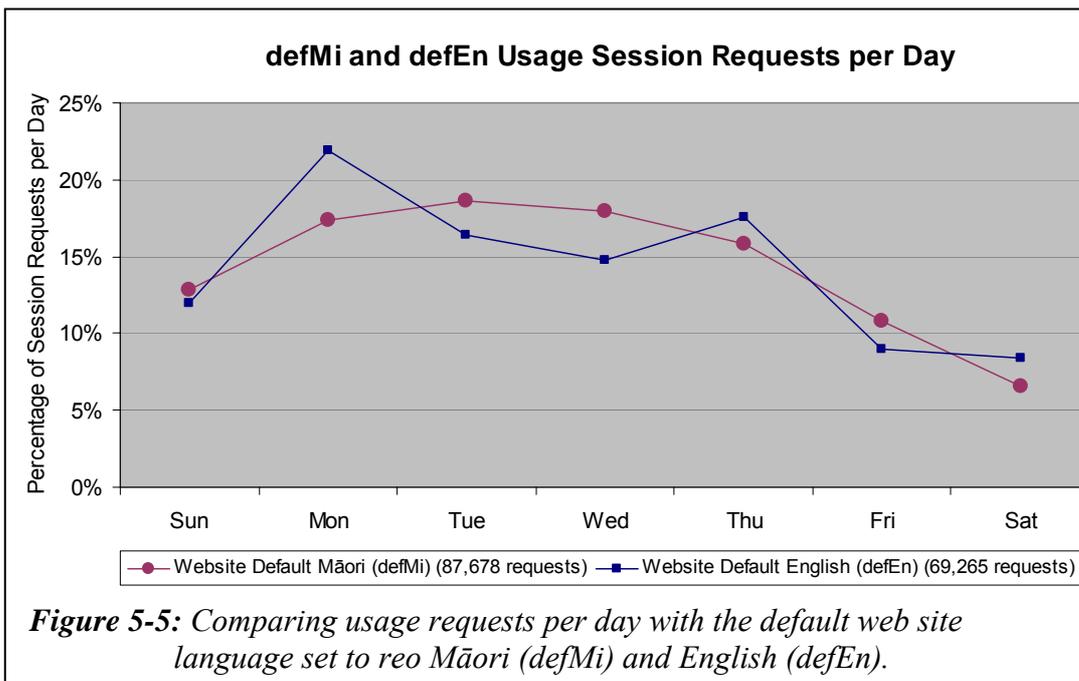


Figure 5-5 shows the requests as they occurred each day. Again, it can be seen that the requests occurred in similar time periods irrespective of the default language setting. Most of the activity occurred from Monday to Thursday, with the least activity on a Saturday.



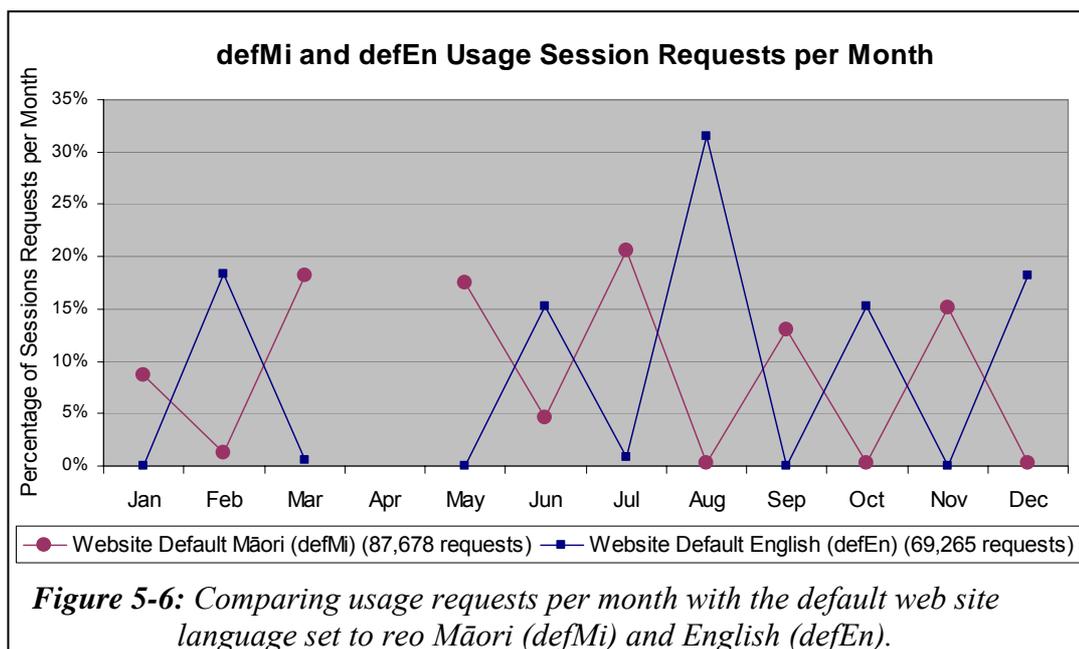


Figure 5-6 shows the requests as they occurred each month. The most requests in te reo Māori corresponded with the months that the default language was defMi. Similarly, months with the defEn setting corresponded with the most requests in English. April shows no requests as they were eliminated from the dataset because of the uncertainty of the default language setting (see §2.2.3). June shows requests from both datasets, indicating that the switching of the default language was late in this month.

The default language setting of the web site did not affect the hours and days of the week that requests were received. Requests recorded each month accurately reflected when the default language setting was alternated.

5.1.6 Summary—Effect on Te Reo Māori Usage

This section has used statistics from TLA to compare the number of reo Māori requests received during a defMi setting with the number received during a defEn setting. The statistics have shown that there was a 6.2% higher reo Māori/English language ratio and 21% more reo Māori usage requests when the default language

was set to te reo Māori. These requests were made by approximately the same number of clients, indicating that the reo Māori default language setting did not encourage new clients to use te reo Māori; rather it encouraged bilingual clients to use te reo Māori more often.

While there was an increase in reo Māori requests there was also a decrease in the number of requests overall. When te reo Māori was used as the Niupepa digital library's default language setting there was a 12.5% reduction in the overall number of requests to the digital library. The proportion of requests in single request, exploratory and usage sessions was the same for both default language settings. This similar proportion implies that the default language setting did not influence the number of exploratory clients who then did not go on to use the Niupepa digital library. Exploratory behaviours of clients are examined in more detail in the next section.

The statistics have also shown that the default language setting of the Niupepa digital library did not appear to influence when the digital library was used by Niupepa clients.

5.2 Effect on Exploratory Behaviours

This section examines how the default language setting affected clients who explored the Niupepa digital library but then did not go on to use it. Although exploratory sessions represented just less than 1% of the total requests, they did give insights into how setting the default language to reo Māori (defMi) or English (defEn) affected the behaviour of exploratory clients.

A high proportion of exploratory sessions were less than two minutes in duration; 80.9% of defMi sessions and 78.1% of defEn sessions. A high proportion of exploratory sessions consisted of just two requests; 83.0% of defMi and 79.7% of defEn sessions, many of which simply involved a request for the home page in one language, then a request for the same page in the other language. With p

values of 0.67 and 0.92 there was no significant statistical difference in the length of defMi and defEn exploratory sessions, either in duration or number of requests.

The defMi and defEn exploratory sessions were then further subdivided by language category. There were:

- 232 requests in 99 defMi expMi sessions at 1.2 requests per day
- 167 requests in 69 defEn expMi sessions at 1.2 requests per day
- 399 requests in 178 defMi expEn sessions at 2.0 requests per day
- 383 requests in 168 defEn expEn sessions at 2.8 requests per day

These figures show that the default language did not have any effect on the number of clients who used te reo Māori to explore (but then not use) the Niupepa digital library. However, it did appear to affect the number of clients who used English to explore the Niupepa digital library; there appeared to be 28.6% fewer expEn requests when the default language of the interface was set to te reo Māori.

5.2.1 Effect on Interface Page Read Times

The lengths of time between requests by expMi clients for interface pages, the apparent read times of interface pages, were calculated. These apparent read times were grouped into read durations of 1–2 seconds, 3–9 seconds, 10–29 seconds, 30–59 seconds, 59–300 seconds and 300+ seconds. The groupings for the four types of exploratory sessions are displayed in *Table 5-1*.

Chi square analysis on the apparent read times revealed some important findings. The p values are summarised in the following list:

- expMi defMi & expMi defEn; 0.85 p value, no significant difference
- expEn defMi & expEn defEn; 0.05 p value, significant difference
- expMi defMi & expEn defMi; 0.0006 p value, significant difference
- expMi defEn & expEn defEn; 0.60 p value, no significant difference

seconds	expMi		expEn	
	defMi	defEn	defMi	defEn
Total	133	98	221	215
1–2	29	23	21	35
3–9	20	13	55	37
10–29	28	25	70	64
30–59	17	15	29	26
59–300	22	13	34	30
300+	17	9	12	23

Table 5-1: defMi and defEn apparent read times of interface pages in expMi and expEn sessions.

The p values mentioned above indicate that clients who explored the Niupepa interface in reo Māori had similar apparent read times whether the default language was set to reo Māori or English. Also, the apparent read times of the expMi clients (for both default language settings) was similar to the apparent read times of the expEn clients when the default language of the web site was set to English. In contrast, the expEn clients exhibited significant statistical differences in apparent read times of interface pages with different settings of the default language.

Statistical summaries were calculated to determine what the differences were in apparent read times of Niupepa interface pages. These summaries are displayed in *Table 5-2*.

expMi and expEn Apparent Read Times for defMi and defEn				
	expMi		expEn	
	defMi	defEn	defMi	defEn
Total Apparent Reads	133	98	221	215
Average (seconds)	171.4	111.9	82.2	109.0
Standard Deviation	485.6	302.8	283.5	269.5
Mode (seconds)	0	0	9	0
Shortest session (seconds)	0	0	0	0
Lower Quartile (seconds)	4	4	8	6
Median (seconds)	21	18	16	18
Upper Quartile (seconds)	84	56	47	59
Longest session (seconds)	3396	3396	3453	1240

Table 5-2: Statistical summaries comparing apparent read times in expMi and expEn sessions with defMi and defEn settings.

High standard deviations indicate a wide variation in the apparent read times. Setting the default interface language to reo Māori corresponded with the greatest extremes in apparent read times. ExpMi clients had the longest average read time (171.4 seconds) when the default interface language was set to reo Māori. ExpEn clients had the shortest read times (82.2 seconds) when the default interface language was set to reo Māori.

Statistics describing the apparent read times of Niupepa interface pages in exploratory sessions reveal some important findings. The expMi page read times were not affected by the default language setting of the Niupepa interface and were similar to the expEn read times when the default language was English. The expEn apparent read times were significantly shorter when the default language was set to reo Māori.

5.2.2 Effect on Which Interface Pages Were Requested

The interface pages requested by exploratory reo Māori (expMi) clients in sessions with a default reo Māori setting (defMi) and in sessions with a default English language setting (defEn) are listed in *Table 5-3*. Also listed are the interface pages requested by exploratory English language (expEn) clients for the defMi and defEn settings of the Niupepa interface. Note that an interface page refers to a page of the Niupepa web site it does not refer to a newspaper document of the Niupepa collection.

Interface Pages Requested in defMi and defEn Exploratory Sessions								
Niupepa Interface Page Requested	expMi				expEn			
	defMi		defEn		defMi		defEn	
Total pages	232	100.0%	95	100.0%	399	100.0%	260	100.0%
niupepa home	192	82.8%	142	85.0%	336	84.2%	310	80.9%
NZDL home	9	3.9%	0	0.0%	30	7.5%	28	7.3%
About Eng. Abstr.	26	11.2%	11	6.6%	13	3.3%	14	3.7%
help	1	0.4%	3	1.8%	6	1.5%	15	3.9%
preferences	4	1.7%	10	6.0%	13	3.3%	15	3.9%
about greenstone	0	0.0%	0	0.0%	0	0.0%	1	0.3%
unknown	0	0.0%	1	0.6%	1	0.3%	0	0.0%

Table 5-3: Numbers and percentages of Niupepa interface pages requested by expMi and expEn clients with defMi and defEn language settings.

With chi square analysisⁱ a comparison of pages requested by expMi clients in defMi and defEn exploratory sessions generates a p value of $3.6 * 10^{-3}$. This indicates that there was a significant difference in the interface pages requested with the different default language settings.

When the default language was set to reo Māori, expMi clients requested more About English Abstracts pages. This could be caused by the ambiguous

ⁱ The requests numbers for ‘about greenstone pages’ and ‘unknown pages’ were not used when undertaking chi square analysis as the counts for these were too low.

translation of this hyperlink term (see §3.3.3 for further explanation). When the default language was set to English there were more requests for the preferences page; a likely scenario given that expMi clients were likely to switch the interface language from English to their preferred language of te reo Māori.

Using chi square analysis to compare the interface pages requested by clients who used English in defMi and defEn exploratory sessions generates a p value of 0.31. This indicates that there was no significant difference in the interface pages requested by expEn users when the default language was set to reo Māori compared with the interface pages requested when the default language was set to English.

5.2.3 Summary—Effect on Exploratory Behaviours

This section examined how the default language setting of the Niupepa interface affected te reo Māori clients who explored the Niupepa digital library but then did not go on to use it. The default language setting did not appear to have any effect on the duration (in time) and number of requests in exploratory sessions, either expMi or expEn. It also did not have any effect on the number of expMi clients who undertook exploratory sessions. However, setting the default language of the interface to reo Māori reduced the number of expEn clients by 28.6%.

The apparent read times of interface pages by expMi clients was similar for both defMi and defEn settings, and these read times were similar to read times taken by expEn clients with a defEn setting. However, apparent read times by expEn clients when the default language was set to reo Māori were significantly shorter.

There were significant differences in the pages requested by expMi clients using defMi and defEn settings. With a defMi setting, the About English Abstracts Page was requested more often; with a defEn setting, the Preferences page was requested more often. There were no significant differences in the Niupepa interface pages requested by expEn clients using the defMi and defEn settings.

5.3 Effect on Interface Behaviours

The previous section examined how the default language of the Niupepa interface affected behaviours by te reo Māori clients who explored the Niupepa digital library but then did not go on to use it. This section examines how the default language of the Niupepa interface affects behaviours by te reo Māori clients who actually did use the Niupepa digital library to request a Niupepa document and/or undertake a query activity.

There were 172,905 usage requests recorded in 6,714 sessions for the Niupepa collection in 2005. Of these, 87,678 (50.7%) requests were recorded in the 198 days that the default language of the web site was set to reo Māori (defMi) and 69,265 (40.1%) requests were recorded in the 137 days that the default language was set to English (defEn). These figures average out to 442.8 defMi requests per day and 505.6 defEn requests per day. The Niupepa digital library received 14% more usage requests when the default language of the web site was set to English. There were almost three times as many clients who preferred to use the interface in English.

Further, and perhaps not surprisingly, the number of usage session requests in a particular language increased when the default language of the web site was set to that language. With a defMi web site setting there were 19,628 (22.4%) requests in te reo Māori and 68,050 (77.6%) requests in English. With a defEn web site setting there were 11,217 (16.2%) requests in te reo Māori and 58,048 (83.8%) requests in English. This difference is illustrated in *Figure 5-3*.

The usage session requests were then further categorised into language session types. The counts for the various defMi and defEn usage session types were:

- usgMi defMi sessions; 92.2 requests/day (18,265 requests in 956 sessions)
- usgMi defEn sessions; 76.5 requests/day (10,475 requests in 623 sessions)
- usgEn defMi sessions; 337.8 requests/day (66,875 requests in 2,222 sessions)
- usgEn defEn sessions; 420.7 requests/day (57,630 requests in 2,322 sessions)

- usgBi defMi sessions; 12.8 requests/day (2,538 requests in 60 sessions)
- usgBi defEn sessions; 8.5 requests/day (3,698 requests in 26 sessions)

These figures show that when the default language of the interface was set to reo Māori, requests by clients who were using the interface in te reo Māori (usgMi) increased by 20.5%. When the default interface was set to English, requests by clients who were using the interface in English (usgEn) increased by 25.4%. With a defMi setting the ratio of usgMi to usgEn clients was 1:2.3; with a defEn setting the ratio of usgMi to usgEn clients was 1:3.8. Clearly, setting the default language of the interface to a particular language encouraged use of the interface in that language.

Bilingual usage sessions (usgBi) were 50.6% more likely to occur if the default language of the interface was set to reo Māori.

Five interface behaviours were further investigated to determine if the default language setting of the Niupepa digital library affected how the interface was used in te reo Māori by Niupepa clients. The interface behaviours investigated were:

- the length of defMi and defEn usgMi sessions
- the apparent read times of interface pages in defMi/defEn usgMi sessions
- the number of return defMi and defEn usgMi sessions
- the interface pages requested in defMi and defEn usgMi sessions
- the interface language switching in defMi and defEn usgMi sessions.

These five interface behaviours are described in the next five sub-sections.

5.3.1 Effect on Usage Session Length

Two measurements are used to define session lengths; duration in time (measured in minutes) and number of requests. Statistics derived from TLA were used to compare lengths of usage sessions when the default language was set to reo Māori and set to English. These comparisons are detailed in the next two sub-sections.

Usage Session Duration in Time

The durations of the usage sessions were grouped into time intervals of less than 1 minute, 1–2 minutes, 3–9 minutes, 10–29 minutes, 30–59 minutes and 60+ minutes. The number of sessions that fitted into these time categories was counted for the different language types of defMi and defEn usage. The counts and the p values of chi square analysis on these counts are displayed in *Table 5-4*.

Duration Groupings of Usage Sessions (in minutes)						
time (minutes)	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
<1	207	224	289	710	0	0
1–2	200	102	523	434	2	2
3–9	185	105	460	352	7	3
10–29	189	103	425	412	22	6
30–59	105	49	275	225	19	10
60+	70	40	250	189	10	5
Total	956	623	2222	2322	60	26
p values	1.5 * 10⁻⁰⁷		9.2 * 10⁻⁴⁴		0.70	

Table 5-4: defMi and defEn comparisons for length of session groupings (in minutes) of usage sessions.

A p value of $1.5 * 10^{-07}$ indicates a significant statistical difference in usage session lengths of usgMi defMi and usgMi defEn sessions. The default language setting affected the duration of usage sessions undertaken by clients who used the Niupepa interface in te reo Māori.

Similarly, with a p value of $9.2 * 10^{-44}$ the default language setting also affected the duration of usage sessions undertaken by clients who used the Niupepa interface in English. By contrast, the length of usage sessions by clients who used the Niupepa interface bilingually was not affected by the default language setting (p value 0.7).

Statistical summaries of session lengths were calculated to determine where the defMi and defEn session duration differences were. These statistical summaries are displayed in *Table 5-5*.

Summary of defMi and defEn Usage Sessions Lengths (Minutes)						
Minutes per session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Average	16.8	15.5	22.9	17.9	44.1	42.7
Standard Deviation	27.3	34.7	36.6	34.8	50.2	42.3
Most Common	0	0	0	0	21	21
Shortest Session	0	0	0	0	3	3
Lower Quartile	0	0	2	0	16	21
Median	5	3	7	4	25	35.5
Upper Quartile	20	14	29	19	48	52
Longest Session	333	321	319	356	256	219

Table 5-5: *defMi and defEn statistical summaries for duration (in minutes) of usgMi, usgEn and usgBi usage sessions.*

The statistics displayed in *Table 5-5* show that the mean, the median and the upper quartile of usgMi sessions were higher with a defMi setting than with a defEn setting. Consequently, usgMi sessions had a longer duration when the default language of the interface was set to reo Māori. Similarly, usgEn sessions also had a longer duration with a defMi setting.

Usage Sessions Length in Number of Requests

Usage session length analysis was also undertaken on the number of requests submitted per usage session. Requests were collated into request counts per session of 2 requests, 3–4 requests, 5–8 requests, 9–19 requests, 20–39 requests and 40+ requests. The counts and the chi square p values are shown in *Table 5-6*.

Requests per session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
2	106	123	107	294	n/a	n/a
3–4	108	94	151	343	n/a	n/a
5–8	247	137	540	478	n/a	n/a
9–19	222	132	545	483	11	8
20–39	147	80	402	337	26	9
40+	126	57	477	387	23	9
Total	956	623	2222	2322	60	26
p values	$2.7 * 10^{-06}$		$1.7 * 10^{-37}$		0.43	

Table 5-6: defMi and defEn request groupings with p values.

A p value of $2.7 * 10^{-06}$ indicates a significant statistical difference in the number of requests in usgMi defMi and usgMi defEn sessions. Consequently the default language setting did affect the number of requests per sessions submitted by clients who used the Niupepa interface in te reo Māori.

A p value of $1.7 * 10^{-37}$ indicates a significant statistical difference in the number of requests in usgEn defMi and usgEn defEn sessions. Statistical summaries were calculated to display the differences in numbers of requests and these summaries are displayed in *table 5-7*.

The Niupepa default language setting affected the number of requests that were submitted by clients who used the interface in te reo Māori. When the default language of the interface was set to reo Māori, more requests were submitted by usgMi clients than when the default language of the interface is set to English. Although this difference did not appear to be much, an average of 19.1 requests compared to an average of 16.8 requests, chi square analysis did show the difference to be statistically significant.

Requests per Session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Average	19.1	16.8	30.1	24.8	42.3	44.6
Standard Deviation	24.5	27.5	42.4	41.8	30.7	45.9
Mode	2	2	6	2	21	30
Fewest requests	2	2	2	2	10	14
Lower Quartile	5	3	6	4	22	19
Median	10	7	15	10	30.5	30
Upper Quartile	23	18	35	28	53	53
Most requests	218	245	598	463	168	239

Table 5-7: *defMi and defEn comparisons of statistical summaries for length (in requests) of usgMi, usgEn and usgBi usage sessions.*

Clients who used the interface in English also submitted significantly more requests when the default interface language was set to reo Māori. Bilingual clients did not show a significant difference in the number of requests submitted per session for the two default language settings of the Niupepa web site.

5.3.2 Effect on Time to Read Interface Pages

As mentioned in §2.5.1, an apparent page read time is the length of time from when one page is requested until another page is requested (where a page refers to a page of the interface as opposed to a document from the collection). The apparent read times can be compared for the different default language settings.

Statistical summaries for defMi and defEn apparent read times are listed in *Table 5-8*.

Apparent Read Times of Interface Pages in defMi and defEn Sessions						
Exploratory Sessions:	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Apparent Reads	797	415	2873	2363	304	196
Average (seconds)	44.7	49.1	48.5	45.6	59.9	98.2
Standard Deviation	138.3	151.9	185.8	168.7	178.7	292.8
Mode (seconds)	4	4	5	4	4	3
Shortest (seconds)	0	0	0	0	0	0
Lower Quartile (seconds)	5	5	6	6	6	6
Median (seconds)	10	10	13	12	14	18.5
Upper Quartile (seconds)	29	34	33	31	35	51
Longest (seconds)	1826	2131	2874	3237	2041	2367

Table 5-8: defMi and defEn statistical comparisons of apparent read times of Niupepa interface pages in usgMi, usgEn and usgBi usage sessions.

An initial look at the average read times above would tend to suggest that clients who used the interface in te reo Māori (usgMi) spent less time reading the interface pages when the default language of the web site was set to reo Māori (44.7 seconds compared with 49.1 seconds). Similarly, the figures tend to suggest that usgEn clients spent less time reading the interface pages with a defEn setting (average 45.6 seconds) than with a defMi setting (48.5 seconds).

However a chi square test on the apparent read times by usgMi clients in defMi and defEn sessions generates a p value of 0.75. This value indicates that there was no statistically significant difference in the apparent read times of usgMi clients using the different default language settings of the Niupepa interface. With p values of 0.03 and 0.14 there was also no statistically significant difference in apparent read times of usgEn clients or usgBi clients using the different default interface settings.

The default language setting of the Niupepa interface did not affect how long usgMi clients took to read interface pages of the Niupepa digital library.

5.3.3 Effect on Return Requests

Return requests are requests submitted by clients who have used the Niupepa digital library previously. In 2005 an average of 11.8 return requests were recorded, per day, by usgMi clients using the defMi interface setting. There was also an average of 12.4 return requests recorded, per day, by usgMi clients using the defEn setting. This difference is not significant. The numbers of return clients who used the interface in English did not show a significant difference which could be attributed to the default language setting; there were, on average, 73.4 return requests per day with a defMi setting and 72.5 return requests per day with a defEn setting.

The number of return requests by clients who used the Niupepa digital library in te reo Māori (and English) was not influenced by the default language setting of the Niupepa interface.

5.3.4 Effect on Which Interface Pages Were Requested

The percentages of interface pages requested in usgMi sessions with defMi and defEn settings are compared in *Table 5-9*. Also listed are the percentages of interface pages requested in usgEn and usgBi sessions.

Table 5-9 suggests that altering the default language of the web interface did not affect which interface pages were requested by usgMi clients. The percentages of usgMi pages requested were similar for both the defMi and defEn setting. A chi square test (with the non-requested pages removed) confirms this, with a p value of 0.14. Thus, the default language of the Niupepa digital library did not influence which pages were requested by usgMi clients.

Interface Pages Requested in defMi and defEn Usage Sessions						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total pages	827	432	3,005	2,476	307	202
Niupepa home page	80.5%	81.3%	83.4%	79.6%	84.0%	63.4%
NZDL home page	1.7%	0.5%	1.4%	3.0%	2.9%	0.0%
About English Abstract	6.7%	4.6%	5.6%	5.8%	4.9%	3.5%
help page	3.3%	4.4%	2.1%	2.6%	1.6%	7.4%
preferences page	7.9%	9.3%	7.6%	9.0%	6.5%	25.7%
about greenstone page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
unknown page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
pages per day	4.2	3.2	15.2	18.1	1.6	1.5

Table 5-9: defMi and defEn comparison of percentages of Niupepa web site interface pages requested in usgMi, usgEn and usgBi sessions.

However, the default language of the Niupepa digital library did appear to affect the interface pages requested by usgEn clients and usgBi clients. Both client categories showed different percentages of pages requested for the different default language settings. With p values of 6.7×10^{-5} and 2.4×10^{-11} these differences were statistically significant. The default language of the Niupepa digital library did influence which pages were requested by usgEn and usgBi clients.

5.3.5 Effect on Interface Language Switching

The number of times a client switched the language of an interface per week was recorded and compared for defMi and defEn web interface settings. These numbers are displayed in *Table 5-10*.

defMi and defEn Web Site Interface Language Switching Summaries						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Switches	102	74	753	390	209	124
Māori to English/week	1.6	1.4	23.3	16.1	4.4	3.3
English to Māori/week	2.0	2.4	3.4	3.8	3.0	3.1
Total Switches/week	3.6	3.8	26.7	19.9	7.4	6.4
Switches/Requests	0.56%	0.71%	1.13%	0.68%	8.23%	3.63%

Table 5-10: defMi and defEn total interface switches, switches per week and total switches/total requests for all usage sessions types.

It is apparent from *Table 5-10* that switching the interface language of the Niupepa digital library occurred infrequently. The 176 switches (102 +74) that occurred over all the usgMi sessions account for 0.6% of all usgMi requests. The 1,140 switches (753+390) that occurred in all usgEn sessions account for 0.9% of all requests in usgEn sessions. The 333 (209+124) switches that occurred in all usgBi sessions account for 5.6% of all usgBi requests. However, by definition bilingual sessions must include at least three language switches.

As might be expected, there were higher proportion of reo Māori to English switches when the default language was set to te reo Māori and a higher proportion of English to reo Māori switches when the default language of the interface was set to English. These higher proportions of switches were behaviour displayed by all client types usgMi, usgEn and usgBi clients.

The switching data was also collated by the number of interface language switches that occurred per usage session. This data is displayed in *Table 5-11*. Most of the switches in usgMi and usgEn sessions consisted of just one or two switches per session. Almost half of the usgMi session switching consisted of just one switch per session, with the reminder consisting of two switches per session. By definition the bilingual sessions consisted of only 3+ switches per session.

defMi and defEn Web Site Interface Language Switches per Session						
Switches per session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
1 switch	48.0%	48.6%	75.6%	66.4%	n/a	n/a
2 switches	49.0%	43.2%	21.5%	32.8%	n/a	n/a
3+ switches	2.9%	8.1%	2.9%	0.8%	100.0%	100.0%
	100%	100%	100%	100%	100%	100%
Switch Sessions/ Total Sessions	7.85%	8.67%	29.57%	13.95%	100%	100%

Table 5-11: defMi and defEn switches per session for all usage types.

Did the default language of the Niupepa digital library influence the language switching undertaken by usgMi clients? No, not significantly. While there were minor increases of language switching from reo Māori to English with a defMi setting, and English to reo Māori with a defEn setting, overall these were not statistically significant. *Table 5-10* shows a difference of just 0.2 total switches per week (3.6 compared with 3.8) for the different defMi and defEn settings. *Table 5-11* shows a difference of just 0.82% (7.85% compared with 8.67%) of switch sessions per total sessions for the different defMi and defEn settings. UsgMi clients, unlike usgEn clients, switched the language of the interface no matter what the default language was set to.

5.3.6 Summary—Effect on Interface Behaviours

This section examined how the default language of the Niupepa digital library affected behaviours by te reo Māori clients who used the Niupepa interface. Statistics derived from TLA were used to show how the default language setting affected the number of requests to the Niupepa digital library and five particular aspects of how usgMi clients used the Niupepa interface.

The statistics revealed that there were 21% more requests in te reo Māori when the default language of the interface was set to reo Māori (defMi) and 25% more

requests in English when the default language of the interface was set to English (defEn). With a defMi setting the ratio of usgMi clients to usgEn clients was 1:2.3, with a defEn setting the ratio of usgMi clients to usgEn clients was 1:3.8. Setting the default language of the Niupepa interface to reo Māori resulted in more requests being undertaken in te reo Māori and in a higher ratio of usgMi to usgEn clients.

As mentioned above, statistics from TLA were used to examine five aspects of usgMi client behaviour when using the Niupepa interface. These statistics showed that sessions by usgMi clients had a longer duration and consisted of more requests if the default language of the interface was set to reo Māori. None of the four other aspects examined was affected by the default language setting of the interface. Clients who used the interface in te reo Māori had similar apparent page read times, similar numbers of return users, requested similar interface pages and showed similar language switching behaviours whether the default language of the interface was set to reo Māori or to English.

This summary concludes the examination into how the default language setting of the Niupepa interface affected usage of the interface by usgMi clients. The next aspect to be examined is whether the default language setting affected how te reo Māori clients accessed documents in the Niupepa digital library.

5.4 Effect on Document Access Behaviours

Did the default language affect how usgMi clients accessed documents in the Niupepa digital library? This question is addressed by examining four aspects of document access:

- how the documents were accessed
- which formats were accessed
- which document types were accessed
- the apparent document read times of the documents that were accessed.

5.4.1 Effect on How Documents Were Accessed

Three methods were available to clients accessing Niupepa documents. Documents were able to be selected from a list of results following a query activity, a list of publication series and a list of articles published in a certain date. Statistics from TLA indicated the methods used by usgMi clients when the default language of the interface was set to reo Māori and to English. These statistics are displayed in *Table 5-12*.

How Documents Were Requested With defMi and defEn Settings						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Documents	13,030	7,427	49,533	42,709	1,539	609
From query	77.6%	76.3%	83.9%	78.0%	60.8%	90.1%
From series	14.4%	10.9%	4.3%	9.9%	13.5%	3.6%
From date	6.1%	10.7%	9.6%	9.9%	21.1%	3.4%
Uncertain	1.8%	2.1%	2.2%	2.3%	4.7%	2.8%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Documents per day	65.8	54.2	250.2	311.7	7.8	4.4
p value	3.2×10^{-37}		2.8×10^{-241}		3.5×10^{-39}	

Table 5-12: *defMi and defEn comparison by percentage of how documents were requested by usgMi, usgEn and usgBi clients.*

Using the search facility to submit a query was the most common method of accessing documents. This was true for all client types and for both default interface language settings. This preference was true for both defMi (77.6%) and defEn (76.3%) settings.

However, the p value of 3.2×10^{-37} indicates that usgMi clients were accessing documents in a significantly different manner in each of the default language settings. The significant difference showed in the preference for the series and date lists; when the default language was set to reo Māori usgMi clients displayed a stronger preference for using the series lists (14.4% compared with 6.1%). When

the default language was set to English this preference was less pronounced and usgMi clients displayed an almost equal preference for using the series and date lists (10.9% compared with 10.7%).

The default language of the web site also affected browsing strategies undertaken by usgEn clients. A defMi setting seemed to encourage usgEn clients to use the query facility (83.9% compared with 78.0%) and seemed to discourage the selection of documents from the series list (4.3% compared with 9.9%).

The default language of the web site seems to have had the most effect on the information retrieval strategies of usgBi clients. When the default language was set to reo Māori they submitted a third fewer queries (60.8% compared with 90.1%), the series list was used almost 10% more (13.5% from 3.6%) and the date list was used 17.7% more (21.2% from 3.4%). Setting the default language to reo Māori seemed to encourage the usgBi clients to be more adventurous in their information retrieval strategies.

Thus, the default language setting of the Niupepa web site significantly affected how Niupepa documents were accessed by all client types.

5.4.2 Effect on Document Formats Requested

As mentioned in §3.1, three document formats were available to the client. *Table 5-13* lists the percentages of requests for each of the formats by usgMi clients using both the defMi and defEn settings. Also listed are the percentages requested by usgEn and usgBi clients.

Formats of Requested Documents With defMi and defEn Settings						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Documents	13,030	7,427	49,533	42,709	1,539	609
Extracted Text	65.3%	53.4%	49.8%	62.8%	58.3%	43.0%
Full Size Image	29.8%	40.6%	44.7%	29.1%	36.5%	35.5%
Preview Image	4.1%	5.8%	5.2%	7.3%	4.9%	21.0%
Undefined	0.9%	0.1%	0.3%	0.8%	0.3%	0.5%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
p value	1.9 x 10 ⁻⁷⁴		0		8.0 * 10 ⁻³⁰	

Table 5-13: *defMi and defEn comparisons of document types requested by usgMi, usgEn and usgBi clients.*

The p values listed in *Table 5-13* indicate that the default language setting of the interface had a statistically significant impact on which document formats were requested by clients of the Niupepa digital library. The extracted text format was the favoured format by all client types, with both defMi and defEn settings; however, it seemed to be less favoured when the default language did not match the usage language of the clients.

For usgMi clients the extracted text documents were requested more (65.3% compared with 53.4%) when the default language of the interface was set to reo Māori. When the default language of the interface was set to English the full size images were requested at a higher rate (40.6% compared with 29.8%).

Similarly, the usgEn clients requested a higher percentage (62.85 compared with 49.8%) of extracted text documents when the default language of the interface was set to English. When the default language of the interface was set to reo Māori the full size images were preferred at a higher rate (44.7% compared with 29.1%) by usgEn clients.

5.4.3 Effect on Which Document Types Were Requested

Five different types of documents were available to Niupepa clients. These documents consisted of; newspapers in te reo Māori, newspapers in English, parallel (reo Māori/English) newspapers, abstracts in English and commentary information written in both English and reo Māori (see §3.5.3). The types of documents requested when the default language of the interface was set to reo Māori were compared with those requested when the default language was set to English; significant differences were noted.

Rather than simply counting the document types requested, the document types requested by unique users were counted. Unique user counts give a better picture of what a wider range of people accessed and limits the influence in the statistics of the heavy users. It was also decided to collate only the top 100 unique pages requested, as limiting the analysis to this number gives a better indication of users' initial preferences and, again, reduces the influence in the statistics of heavy users of the web site.

Documents requested by usgBi clients were excluded as the number of requests submitted by usgBi clients were considered too low to generate meaningful results. The data collated for this analysis consisted of:

- 774 unique user requests for 100 pages in usgMi sessions with defMi setting
- 686 unique user requests for 141 pages in usgMi sessions with defEn setting
- 2048 unique user requests for 98 pages in usgEn sessions with defMi setting
- 2185 unique user requests for 110 pages in usgEn session with defEn setting

Requests for each document type are displayed as percentages in *Figure 5-7*.

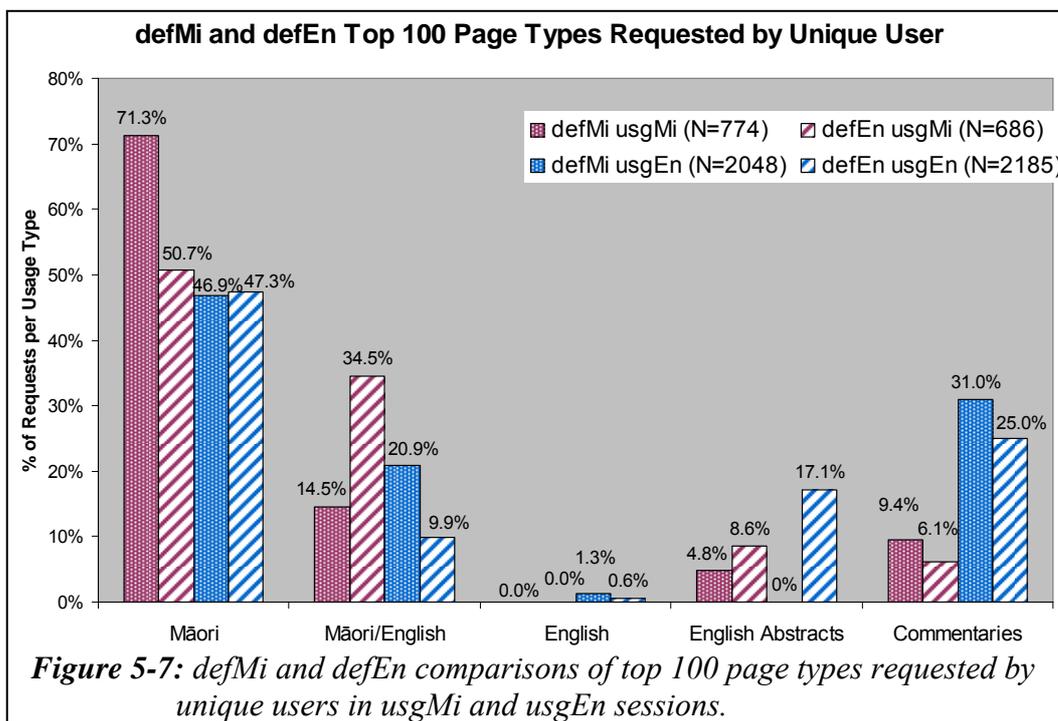


Figure 5-7 shows that the default language of the Niupepa interface did affect the type of documents accessed by usgMi clients. When the default language was set to reo Māori, usgMi clients requested reo Māori newspapers in 71.3% of their requests and parallel (reo Māori/English) newspapers in only 14.5% of requests. However, when the default language was set to English, usgMi clients requested reo Māori newspapers in only 50.7% of their requests and parallel newspapers in 34.5% of requests. UsgMi clients were twice as likely to request English Abstracts (8.6% compared with 4.8%) when the default language was English.

Figure 5-7 also highlights an unusual behaviour by usgEn clients. The English Abstracts were requested in 17.1% of requests when the default language of the web site was set to English, but, incredibly, weren't requested at all when the default language of the web site was set to reo Māori. It is difficult to imagine why a usgEn client would not want to select an English Abstract simply because the default language of the web site was set to reo Māori. Perhaps with the interface initially set to reo Māori, the usgEn clients were not aware that English Abstracts were available.

5.4.4 Effect on Reading Time Of Documents

The apparent document read times (as explained in §2.5.1) by usgMi clients were compared for the defMi and defEn settings. The average apparent read times of full size documents, of extracted text documents and of preview documents were calculated and are displayed in *Table 5-14*.

Average Apparent read times (seconds)	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total	70.9	73.9	56.2	53.0	77.8	57.9
Full Size Image	117.7	106.1	90.7	96.6	135.3	86.1
Extracted Text	52.3	54.9	29.1	36.9	47.0	53.1
Preview Image	43.4	23.0	26.7	25.5	27.9	22.2

Table 5-14: *defMi and defEn comparisons of average apparent read times (in seconds) of documents by usgMi, usgEn and usgBi clients.*

These figures indicate that when the default language was set to reo Māori (defMi) the usgMi clients took longer, on average, to read the full size images (117.7 seconds compared with 106.1 seconds) and the preview images (43.4 seconds compared with 23.0 seconds). However, there was no significance difference in the apparent document read times of extracted text. The default language setting did affect the apparent read times of preview and full size images but not extracted texts.

5.4.5 Summary—Effect on Document Access

This section examined how the default language setting of the Niupepa interface affected how te reo Māori clients accessed documents. Statistics produced from TLA were used to reveal significant differences in four document access behaviours.

Using a query was the method preferred by Niupepa clients to access documents. Statistics showed that 75% of the usgMi requests for documents resulted from a query, a rate consistent for both default interface language settings. However, usgMi clients showed a higher preference for using the series list when the default language was set to reo Māori than they did when the default was set to English.

Extracted text was the preferred document format in 56% of Niupepa requests. There were 12% more requests by usgMi clients for the extracted text format with a defMi setting. Alternatively, there were 11% more requests by usgMi clients for the full size image format with a defEn setting.

The default language setting of the Niupepa interface also affected the type of documents requested by usgMi clients. UsgMi clients requested significantly more reo Māori newspapers and significantly fewer parallel (reo Māori/English) newspapers when the default interface language was set to reo Māori. UsgMi clients also requested more English Abstracts when the default interface language was set to English.

The apparent read times of extracted text documents were similar for both the defMi and defEn settings. However, usgMi clients took longer to read full size images and preview images when the default language was set to reo Māori.

This section analysed how the Niupepa default language setting affected document access behaviours of te reo Māori clients. The third aspect of behavioural analysis, query behaviours, is examined in the next section.

5.5 Effect on Query Behaviours

This section analyses how the default language setting affected the query behaviours of clients who used the Niupepa digital library in te reo Māori. A query is defined as using the search facility of the Niupepa digital library to find a term or string of terms in documents stored in the Niupepa digital library.

The TLA process generated query behaviour statistics on the data collected when the Niupepa default language was set to reo Māori (defMi) and to English (defEn). Five query behaviours of usgMi clients are examined in this section:

- the number of queries submitted
- the number of query terms used
- the language of query terms
- the use of query parameters
- the use of query operators.

5.5.1 Effect on Number of Queries

The numbers of queries submitted per session by usgMi clients when the default language was set to reo Māori were compared with those submitted when the default was set to English. Percentages were calculated and are listed in *Table 5-15* for comparison. Also displayed are the percentages of queries per session by usgEn and usgBi clients for the defMi and defEn settings.

Table 5-15 shows that the default language setting of the Niupepa interface did affect the number of queries per usgMi session. UsgMi sessions were more likely to involve a query when the Niupepa interface was set to defMi.

defMi and defEn Sessions That Contained Queries						
	usgMi		usgEn		usgBi	
Queries/Session	defMi	defEn	defMi	defEn	defMi	defEn
no queries	39.4%	50.6%	27.0%	41.1%	20.0%	23.1%
1–2 queries	32.5%	22.5%	35.0%	28.6%	13.3%	11.5%
3+ queries	28.0%	27.0%	38.0%	30.3%	66.7%	65.4%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Total Sessions	956	623	2222	2322	60	26

Table 5-15: defMi and defEn comparisons of number of sessions that contained queries by usgMi, usgEn and usgBi clients.

It appears that approximately 11% of the usgMi sessions were influenced by the defMi and defEn setting. With a defMi setting 60.5% (32.5% + 28.0%) of the usgMi sessions involved the submission of at least one query, with a defEn setting, only 49.5% (22.5% + 27.0%) of the usgMi sessions involved the submission of at least one query.

The default language setting also affected the number of queries that were submitted per usgEn and per usgBi session. Sessions undertaken by usgEn and usgBi clients showed a significantly higher rate of queries per session when the default language of the interface was set to reo Māori. Both usgEn and usgBi clients had a higher percentage of queries submitted per sessions than usgMi clients.

5.5.2 Effect on Number of Query Terms

The numbers of query terms in queries submitted by usgMi clients with a defMi setting were compared with those submitted by usgMi clients with a defEn setting. The percentages of these terms were calculated and are displayed in *Table 5-16*. Also displayed are the numbers of usgEn and usgBi query terms for the different default language settings of the Niupepa interface.

	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
1 term/query	45.6%	48.1%	47.8%	51.2%	43.9%	52.5%
2 terms/query	28.8%	31.9%	33.3%	30.8%	39.4%	24.6%
3+ terms/Query	25.7%	20.0%	18.9%	17.9%	16.6%	23.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Queries	2666	1703	8650	6871	289	183
p value	$8.1 * 10^{-5}$		$1.0 * 10^{-4}$		$3.3 * 10^{-3}$	

Table 5-16: defMi and defEn percentage comparisons for the number of terms in a query in usgMi, usgEn and usgBi sessions.

The default language setting affected the number of query terms used per query by usgMi clients. With a p value of $8.1 * 10^{-5}$ the usgMi sessions showed a statistically significant difference in the number of query terms used for the defMi and defEn settings. The percentage of usgMi queries with just one term was lower with the defMi setting (45.6% compared with 48.1%). The percentage of usgMi queries with 3+ query terms was higher for the defMi setting (25.7% compared with 20.0%). These percentages indicate that with a defMi setting usgMi clients submitted more query terms per query.

The default language setting also affected the number of query terms used per query by usgEn and usgBi clients; they also entered more query terms per query when the default language of the interface was te reo Māori.

5.5.3 Effect on Language of Query Terms

The language of the query terms most requested by usgMi clients were examined to see if the default language setting had an influence over which language was preferred. These consisted of 116 terms submitted 479 times when the default language was set to te reo Māori and 85 terms submitted 352 times when the default language was set to English. The percentages of query terms in te reo Māori and in English are displayed in *Table 5-17*.

The default language setting of the Niupepa interface did affect the language of query terms used by usgMi clients. With a defMi setting, 94.4% of their query terms were entered in te reo Māori. With a defEn setting only 86.1% of their query terms were entered in te reo Māori. Setting the default language to te reo Māori corresponded with a higher number of query terms submitted in te reo Māori.

Language of Query Terms With defMi and defEn Settings						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
reo Māori terms	94.4%	86.1%	83.3%	83.0%	74.8%	74.7%
English terms	5.6%	13.9%	16.7%	17.0%	25.2%	25.3%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Terms Examined	479	352	965	962	163	83
Unique Terms	116	85	104	118	62	30

Table 5-17: defMi and defEn percentage comparisons for language of query terms submitted by usgMi, usgEn and usgBi clients.

Interestingly, the default language of the Niupepa interface did *not* affect the language of the query terms used by usgEn and usgBi clients. UsgEn clients showed similar percentages of query terms in te reo Māori for both default language settings (83.8% compared with 83.0%). UsgBi clients also showed similar percentages of query terms in te reo Māori for both default language settings (74.8% compared with 74.7%).

5.5.4 Effect on Altering of Query Parameters

Earlier statistics (see §3.6.5) indicated that the only query parameter to be altered significantly by Niupepa clients was the Some/All parameter. With a defMi setting 7.0% of usgMi clients altered this parameter, while with a defEn setting 9.7% altered this parameter. This difference was not significant enough to conclude that the default language affected how much this query parameter was used by usgMi clients.

The default language setting did seem to affect how often the Some/All query parameter was used by usgEn clients. With a defMi setting, 12.1% of usgEn clients altered this parameter, while with a defEn setting only 8.8% altered this

parameter. It appears that if usgEn clients were prepared to switch the language of the interface from reo Māori to English they were also more likely to alter the Some/All query parameter.

5.5.5 Effect on Use of Query Operators

Earlier statistics (see §3.6.6) indicated that the only query operators used significantly by Niupepa clients were quotation marks; those were used to define precise phrase querying. With a defMi setting, 16.6% of queries submitted by usgMi clients included the use of quotation marks, while with a defEn setting 17.1% of queries included quotation marks. Thus, the default language setting of the Niupepa web site did not affect the number of query operators used by usgMi clients.

The default language setting did appear to affect the number of usgEn clients who used the quotation marks to define precise phrase querying. With a defMi setting, 4.9% of queries submitted by usgEn clients included the use of quotation marks. This changed to 8.9% of queries with a defEn setting. UsgMi clients were only half as likely to use quotation marks if the default language of the interface was set to reo Māori.

5.5.6 Summary—Effect on Query Behaviours

This section analysed how the default language setting affected the query behaviours of clients who used the Niupepa digital library in te reo Māori. Statistics produced from TLA were used to examine five query behaviours of usgMi clients and identify differences arising from the different default language settings of the Niupepa interface.

The statistics revealed that the default language setting of the Niupepa interface significantly affected three query behaviours of usgMi clients. When the default

language of the interface was set to reo Māori usgMi clients were more likely to submit a query, used more query terms per query and were more likely to enter the terms in te reo Māori.

The default language setting of the Niupepa interface did not appear to affect how often the Some/All query parameter was used by usgMi clients or the use of quotation marks to define precise phrase matching.

Analysis in the last three sections of this chapter has revealed that setting the default language to reo Māori encouraged the use of the Niupepa digital library in te reo Māori. With a defMi setting there was more use of the interface in te reo Māori, more te reo Māori terms when querying and documents written in te reo Māori were requested more. Before this chapter is concluded a further analysis merits discussion—how the default language setting affected clients who used the Niupepa digital library in English. This question is addressed in the following section.

5.6 Effect on Use in English

The previous three sections have examined how the default language setting affected the usage of the Niupepa digital library by clients who predominantly used the interface in te reo Māori (usgMi). Those sections showed that use of te reo Māori in almost all aspects of information retrieval increased when the default language of the interface was set to reo Māori. However, given that there were almost three times as many usgEn as usgMi clients, a further question needs to be considered. How did setting the default language to reo Māori affect usage of the Niupepa digital library by clients who predominately used the interface in English?

Setting the default language to reo Māori reduced the total number of filtered requests. Statistics from TLA revealed that 12.5% fewer Niupepa requests were received with a defMi setting. The number of requests by usgEn clients was

reduced even more as 6.2% of the usgEn requests had shifted to be recorded as usgMi requests.

While the number of usgMi clients remained constant for both default language settings, this was not the case for the usgEn clients. There were 1/3 (36.2%) fewer usgEn clients when the default language of the interface was set to reo Māori.

With a defMi setting, usgEn clients undertook longer usage sessions (both in duration and number of requests), requested the Niupepa home page at a higher rate and switched the interface language to English more often. These behaviours were perhaps a consequence of the default language setting not matching the clients' preferred language of use and, thus, these clients had to undertake the extra step of switching the interface language into a language that they could use.

When the default language was set to reo Māori, usgEn clients were more reliant on the query facility to access documents and had a higher preference for requesting full size images. One surprising aspect was that usgEn clients did not request English Abstracts when the default language of the interface was set to reo Māori. This behaviour is perhaps a usability awareness issue with the Niupepa web site as English Abstracts were created as a resource for usgEn clients no matter what the setting of the default language was.

UsgEn clients submitted more queries per session during a defMi setting. They also altered the Some/All query parameter more often and submitted more query terms per query with this setting. A higher number of query terms per query often means more use of the quotation marks to define phrase matching; however, this operator was not used by usgEn clients as much with a defMi setting.

In general, when the default language was set to reo Māori, usgEn clients used fewer features of the Niupepa interface and were more reliant on the query facility to access documents.

It is difficult to confirm that the behavioural differences noted by the usgEn clients with the different default language settings were solely a result of the default language setting of the Niupepa interface. The behaviours displayed by the

usgEn clients with a defMi setting could be the behaviours that these clients undertook for both default language settings. However, these behaviours only become apparent when the 1/3 of usgEn clients intolerant of the defMi setting were not included in the statistics.

What can be confirmed is that 1/3 of the usgEn clients were unwilling to use the Niupepa digital library if the default language was set to reo Māori, resulting in 12.5% fewer requests to the Niupepa digital library when the interface was set to te reo Māori.

5.7 Summary

This chapter has continued the investigation into how indigenous languages are used by clients of digital libraries. In particular, this chapter examined how the default language setting of the Niupepa digital library affected te reo Māori usage of the Niupepa digital library. The results have shown that setting the default language to reo Māori significantly increased the number of usgMi requests and, perhaps more importantly, increased the number of situations where te reo Māori was used to undertake information retrieval activities.

Setting the default language to reo Māori led to a 21% increase in reo Māori usage requests to the Niupepa digital library. This increase was mostly from clients who were already using the digital library in te reo Māori.

Setting the default language to reo Māori had very little effect on how the Niupepa digital library was explored (but then not used) by clients who used te reo Māori. The number of te reo Māori exploratory clients, the duration of te reo exploratory sessions, the number of requests in reo Māori exploratory sessions and the apparent interface page read times in reo Māori exploratory sessions were all similar for both default language settings. The only difference noted was the number of reo Māori exploratory clients who accessed the About English

Abstracts button; a difference most likely caused by an ambiguity issue arising from an inaccurate translation of the button label.

Setting the Niupepa default language to reo Māori did affect the length of reo Māori usage sessions; they were of a longer duration and consisted of more requests. However, other usgMi interface characteristics (such as apparent page read times, numbers of return users, interface pages requested and language switching behaviours) were unaffected by whether the default language was set to reo Māori or English.

The manner in which te reo Māori clients accessed documents was affected by the default language setting. These clients showed a stronger preference for using the series list, accessing extracted text formats, accessing Niupepa documents written in te reo Māori, and took longer to read full size and preview images when the default language was set to reo Māori.

Some query behaviours of te reo Māori clients were affected by the default language setting of the Niupepa web site. UsgMi clients were more likely to submit a query, used more query terms per query and were more likely to enter the terms in te reo Māori when the default language of the interface was set to reo Māori. However, use of the Some/All query parameter and the quotation marks query operator were not affected by the default language setting of the Niupepa web site.

The default language setting of the Niupepa interface also affected clients who used the interface in English. Setting the default language of the interface to reo Māori reduced the number of English exploratory (expEn) requests by 28.6% and the number of English usage (usgEn) requests by 19.7%. When the default language of the interface was set to reo Māori clients who used the digital library in English used fewer features of the Niupepa interface and were more reliant on the query facility to access documents.

5.7.1 Discussion

Data from the 2006 census (Statistics New Zealand, 2006) indicates that 98.5% of Māori are able to speak English. Developers of a reo Māori/English bilingual information system could, perhaps, assume that English would be a suitable default language as it would be useable for both English speaking and reo Māori bilingual clients. However, setting the default language of a bilingual system to English would not assist in the revitalisation of te reo Māori. In fact, it would do the opposite. Statistics derived from TLA on the Niupepa digital library in 2005 have shown te reo Māori was used significantly less by te reo Māori clients when the default language was not set to reo Māori.

Alternatively, setting the default language of a bilingual interface to an indigenous language encourages use of the interface in the indigenous language. Setting the default language of the Niupepa digital library to reo Māori not only resulted in significantly more requests with the interface language set to reo Māori but also resulted in te reo Māori being used in more situations to undertake information retrieval. Examples of these situations included a higher use of the series list (where te reo Māori is prominent), a higher number of reo Māori documents accessed, longer read times of those documents and a higher percentage of query terms written in te reo Māori.

This chapter has shown that the assistance of a bilingual digital library to the usage of an indigenous language can increase in two areas if the default language of the digital library is set to the indigenous language. First, the amount of times that the indigenous language is displayed and, subsequently, used by clients is significantly increased. Second, the new opportunities and avenues available for using the indigenous language are exploited more often if the default language is set to the indigenous language.

6 Hawaiian Language Usage of the Nūpepa Digital Library

Chapter 6 discusses how the Hawaiian language was used in 2005 by clients of the Hawaiian Nūpepa digital library—Nā Nūpepa ‘Ōlelo Hawai‘i. Hereafter, this digital library will be referred to simply as the Nūpepa digital library and should not be confused with the Niupepa digital library which refers to the Aotearoa Māori Niupepa digital library.

The Nūpepa digital library was another digital library suitable for analysing the use of an indigenous language in a modern electronic medium. Like its Māori counterpart, it provided a significant quantity of indigenous language resources, through an indigenous and non-indigenous switchable interface and was freely available on the Internet.

Was the Hawaiian language used by clients of the Nūpepa digital library to access documents written in Hawaiian? Results from TLA undertaken on the Nūpepa dataset indicate that the Hawaiian language was the preferred language in over 200,000 requests to the Nūpepa digital library in 2005. Given that the Hawaiian language was used by Nūpepa clients, the subsequent question to pose is how was it used?

This chapter details how the Hawaiian language was used in the Nūpepa digital library in 2005. The chapter begins with an introduction of the Hawaiian Nūpepa digital library, where the content was sourced, how the digital library was established and what differences there were between the Hawaiian Nūpepa digital library interface and the Māori Niupepa digital library interface. This is followed

by a section which details how much the Hawaiian language was used in the Nūpepa digital library in 2005.

Exploratory session data is analysed to determine how Hawaiian language clients explored the Nūpepa digital library but then did not go on to use it. This data reveals that Hawaiian exploratory sessions on average lasted less than three minutes and consisted of three requests or fewer.

Usage session data will be analysed in three sections; how Hawaiian language clients used the Nūpepa interface, how they accessed Nūpepa documents and submitted queries. The three sections provide an extensive examination of how the Hawaiian language was used by Nūpepa clients.

Two further topics are examined using statistics generated from TLA. First, Hawaiian language usage of the Nūpepa digital library is compared to English language usage. This will reveal that Hawaiian language clients undertook longer sessions, used different methods to access documents and preferred different document formats from English language clients. The second and final topic examined is the use of the Nūpepa digital library by clients who preferred to access documents bilingually, that is, using both Hawaiian and English. A quarter of all requests were recorded in sessions where clients used the Nūpepa interface bilingually. Bilingual clients exhibited unique querying behaviours that led to a high number of documents accessed per query.

Chapter 6 concludes by summarising how the Hawaiian language was used by clients of the Nūpepa digital library.

6.1 The Hawaiian Nūpepa Digital Library

The next four sub-sections give introductory information about the Hawaiian Nūpepa digital library. A brief history is given about the content of the Nūpepa digital library, newspapers written in Hawaiian, and how these were captured and placed in an online environment. This is followed by a description of the Nūpepa interface, which includes screenshots taken from the Nūpepa web site. This section concludes by detailing the differences between the Nūpepa and Niupepa interfaces.

6.1.1 Capturing the Hawaiian Nūpepa

The publication of newspapers in Hawai‘i began early in its colonial history. The first two newspapers, *Ka Lama* (The Torch) and *Ke Kumu* (The Teacher), both published in 1834, were Protestant mission-sponsored newspapers. These publications began a myriad of newspaper publications that occurred in the Hawaiian Islands. Some estimates suggest over 1,000 separately-titled newspapers have been published there in at least nine different languages (Chapin, 1996), of which just over 100 were in the Hawaiian language. Although the newspapers were diverse, Chapin categorises them into four main groupings: establishment, opposition, official and independence. The amount of information and the unique viewpoints ensure that the newspapers are rich and valuable sources of historical information.

In the mid-1980s, efforts began in the capture and preservation of the newspaper pages that had survived until then. Although the large majority of newspaper publications had not been preserved, there were still significant numbers of newspaper pages that had been stored in the various libraries, archival collections and museums throughout Hawai‘i. A grant was awarded to the University of Hawai‘i at Manoa, by the National Endowment for the Humanities, to copy all remaining historic Hawaiian and ethnic newspapers onto microfilm. In total 260,000 pages were microfilmed, including papers in Hawaiian, English and

Portuguese. While there were names for 103 known Hawaiian language newspapers, only about half of these had surviving hard copies that were available to be microfilmed. These hard copies comprised about 60,000 discrete newspaper pages.

The Hawaiian newspapers are written predominantly in the Hawaiian language. It was a surprise to researchers recovering the images to discover that some of the pages were, in fact, written in English. These writings are small in number, sometimes a column, sometimes a page, and are scattered throughout some of the newspaper pages in a small number of the periodicals. Although there are no exact figures it appears the English writings comprise an insignificant proportion of the writings in the Hawaiian language newspapers.

6.1.2 Hawaiian Nūpepa Online

Work on placing the Hawaiian language newspapers online began in 1997. Some of the microfilm copies of the newspapers had begun to deteriorate as a result of frequent use by students and scholars of the Hawaiian language. Three libraries of the University of Hawai'i received a small grant to undertake a pilot project to digitise a small number of the heavily used newspapers pages and place them on the web. Further small grants were received in 1998, 1999 and a more substantial grant in 2000, which allowed more titles and more newspaper pages to be added to the online collection (Cartwright, Chantiny, Hori, & Peacock, 2000). At the completion of this funding approximately 2,000 pages of Hawaiian language newspapers could be viewed online.

Having the Hawaiian language newspapers online created much interest throughout the Hawaiian language communities, and with researchers and scholars of Hawaiian language and history. Investigations began into the possibility of placing all Hawaiian language newspaper pages online and making them full text searchable. Initial discussions involved Dwayne Nakila Steele, a philanthropist and retired businessman, Puakea Noglemeier, professor of Hawaiian Language at the University of Hawai'i at Manoa, the Hawaiian

Language Legacy Programme of Alu Like Inc.ⁱ, and Hale Kuamo‘oⁱⁱ. Alu Like Inc. had at that time been working with Hale Kuamo‘o to place a journal of Hawaiian literature online, so had already established a working relationship and had begun looking at the issues of hosting a digital library web site. Steele provided personal funding and was able to secure funding from other private sources.

Capturing the text from the Hawaiian newspaper pages proved to be difficult for three reasons. First, the volume of the text per page was very large; tabloid (597mm × 375mm) and broadsheet (749mm × 597mm) page sizes were common in the newspapers many of which can consist of up to six columns. On average there were over 6,300 words per newspaper page. Second, the quality of the captured image was often poor; smudging, fading, skewed text or gutter/baseline defects were often present in the newspaper pages, all of which reduce the ability of the OCR process to accurately generate texts. Third, there was very little computer software support for the Hawaiian language; consequently the OCR program had to be reconfigured and dictionaries had to be created to enable the OCR process to read the Hawaiian language.

The extracted text was generated automatically and, in the early stages of the web site development, there was minimal work done on correcting the errors resulting from this process. However, after much debate it was decided that the error rate was unacceptable and that Hawaiian language experts were needed to carefully correct the texts generated from the OCR process. In 2003, the Ho‘olaupa‘i

ⁱ Alu Like Inc. is the largest non-government, non-profit service organisation that assists native Hawaiians to achieve social and economic self sufficiency. Alu Like Inc. has services which include economic and business development, employment preparation and training, library services, education and childcare services. Alu Like Inc. is funded by federal, state, county and private sources and is governed by a state-wide volunteer Board of Directors.

ⁱⁱ Hale Kuamo‘o is the the Hawaiian Language Center within Ka Haka ‘Ula O Ke‘elikōlani (the Hawaiian Language College) of the University of Hawai‘i at Hilo. Hale Kuamo‘o was established by Hawai‘i State Legislature in 1989 to encourage, expand and sustain Hawaiian language, Hawaiian culture and Hawaiian-medium education in all situations. Its activities include providing publications and training for Hawaiian-medium teachers, making available Hawaiian literature for various forms of media and undertaking research into Hawaiian dictionaries and grammar terminology.

project was established at the Bishop Museum in Honolulu to partner with Alu Like Inc. and Hale Kuamo‘o and undertake OCR of the Hawaiian Newspapers to a high standard of accuracy (upwards of 99.8%). Hawaiian language students were employed to undertake the laborious task of OCR and Hawaiian language experts were used to verify the extracted texts. However, because of funding constraints, the difficulties listed above and the level of accuracy required, progress on getting all the Hawaiian language pages converted into an accurate text has not been rapid. By early 2007 approximately 100 to 150 extracted text pages were being generated per month.

The newspapers are located on a web site called Ulukau—the Hawaiian Electronic Library web site. As mentioned in §2.2 Ulukau is also the name of the web server that makes available the Nūpepa collection. The Ulukau web site can be found online at <http://ulukau.org>. A private company called DL Consulting Ltd, located in Hamilton, Aotearoa, was contracted to manage the process of making these and other Hawaiian language resources available via the Internet. DL Consulting built the Ulukau web site using the open source digital library software called Greenstone (Greenstone software is described in §1.2.3). The purpose of the Ulukau web site is to make Hawaiian language resources available for the use, teaching and enhancement of the Hawaiian language.

The Ulukau web site was established in 2004 and operates through the co-operation of dozens of agencies and individuals. Its leading two partners are Hale Kuamo‘o and the Native Hawaiian Library. The Native Hawaiian Library is one of the major programmes funded by Alu Like Inc., Inc. It provides library and information services to native Hawaiians and others who are interested in Hawaiian information and documents. The Ulukau web site is an ongoing web site with more documents and more collections being added as they become available.

As at June 2006 there were seven electronic collections publicly available at this web site:

- Puke Pai ‘Ōlelo Hawai‘i—Journal Of Hawaiian Language Sources
- Nā Palapala o ka Māhele—Māhele (land) Database
- Nā Nūpepa ‘Ōlelo Hawai‘i —Hawaiian Newspapers
- Nā Inoa ‘Āina—Hawaiian Place Names
- Ka ‘Ohina ki‘i Na Ed Greevy—Ed Greevy’s Photograph Collection
- Nā Puke Wehewehe ‘Ōlelo Hawai‘i —Hawaiian Dictionaries
- Nā Puke ‘Ē A‘e—More Books

6.1.3 Hawaiian Nūpepa Interface Description

The Hawaiian language newspapers, located on the Ulukau web site can be accessed directly via the URL www.nupepa.org. By July 2006 there were 53,425 Hawaiian language newspaper pages available on this web site. According to the Hawaiian Language Legacy Programme (Stauffer, 2006), these pages constitute nearly all of the surviving Hawaiian language newspaper pages. These are from just over 100 reels of microfilm and represent 46 different newspapers. The fewest number of surviving pages from a newspaper was two pages from *Ko Hawai‘i Pae Aina*, while the largest number of surviving pages was 18,153 pages from the periodical *Ka Nupepa Kuokoa*.

No exact records were kept for the year of 2005; however, a calculated estimate would suggest that in 2005 there were an average of 2,550 text files available out of a total of 40,000 discrete documents, representing 6.4% of the total documents.

The Nūpepa web site is a bilingual web site that, by default is displayed in the Hawaiian language. There is a button located in the top right-hand corner that can alternate the interface language display between English and Hawaiian. This button is available on all pages of the Nūpepa web site. The Nūpepa home page gives some explanatory information about the Nūpepa Collection and makes available three methods to access the newspapers; search (*huli*), titles (*inoa nūpepa*) and dates (*lā ho‘opuka*). The Nūpepa web site home page is displayed in *Figure 6-1*.

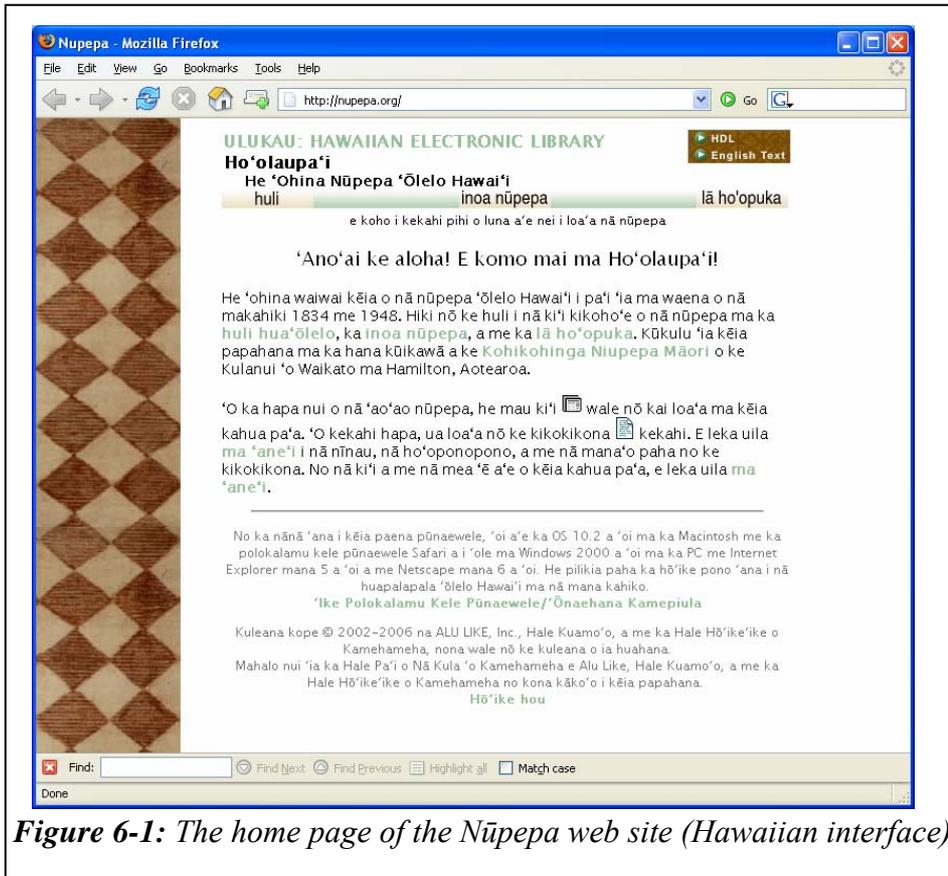


Figure 6-1: The home page of the Nūpepa web site (Hawaiian interface).

The first method available to access Nūpepa documents is by undertaking a query (huli). The English version of the query page is displayed in *Figure 6-2* and shows the results of a query for the term *'nupepa'*. It can be seen that the query interface screen is simple, with just the one query argument able to be altered; the Some/All option. Clicking on an icon underneath the results bar results in either a graphic page or an extracted text document being displayed (as in *Figure 6-3*).

When the extracted text page is displayed, the query term is highlighted as can be seen in *Figure 6-3*. This page provides links to access the previous newspaper page, the next newspaper page, or any page in the current edition of the newspaper. There are also links to a preview image, a mid-size image, a full size image, a link to detach the text to a separate browser window and a link that will display the text without any highlighting.

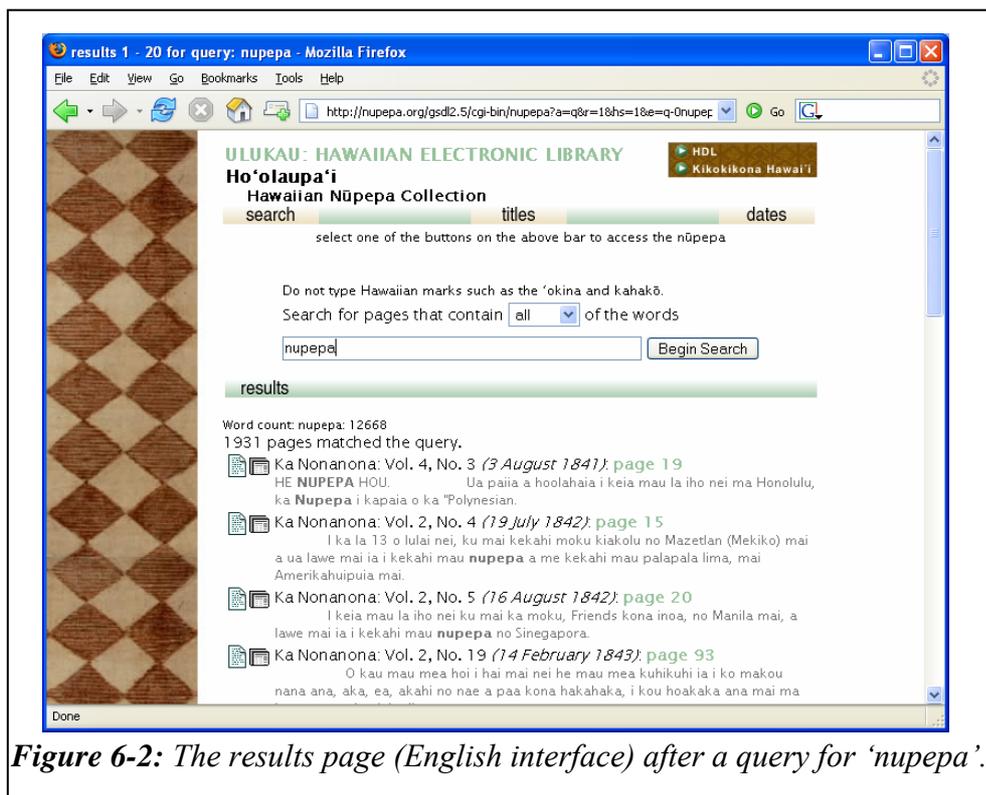


Figure 6-2: The results page (English interface) after a query for 'nupepa'.

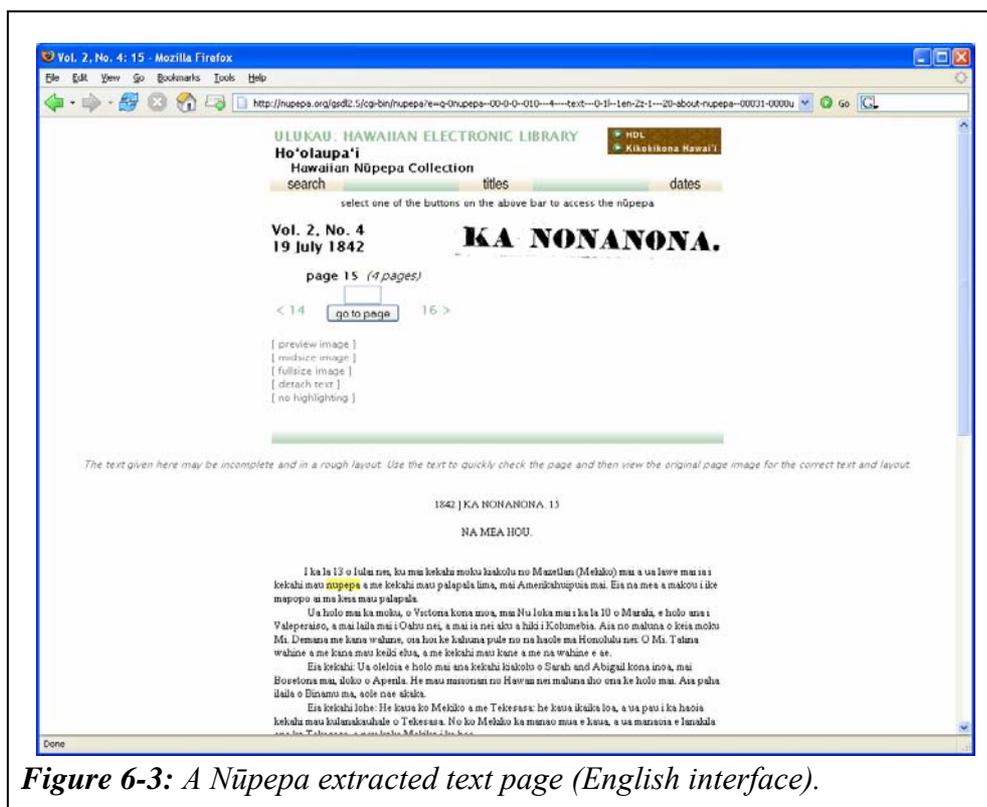


Figure 6-3: A Nūpepa extracted text page (English interface).

The second method of accessing Nūpepa documents is to select a document from the series list. Clicking on the titles (*inoa nūpepa*) button will display a series list of titles as shown in *Figure 6-4*.

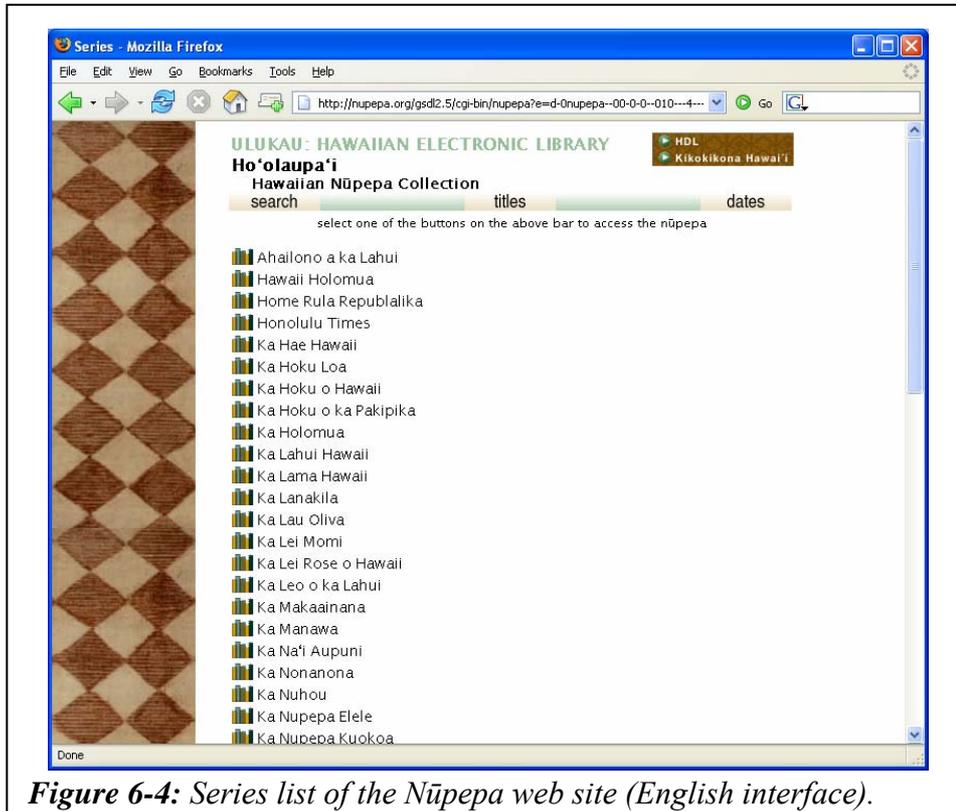


Figure 6-4: Series list of the Nūpepa web site (English interface).

The third and final method of accessing Nūpepa documents is to select a document from a grouping by date. Clicking on the dates (*lā ho'opuka*) button and then the year 1862 will lead to the screen displayed in *Figure 6-5*.

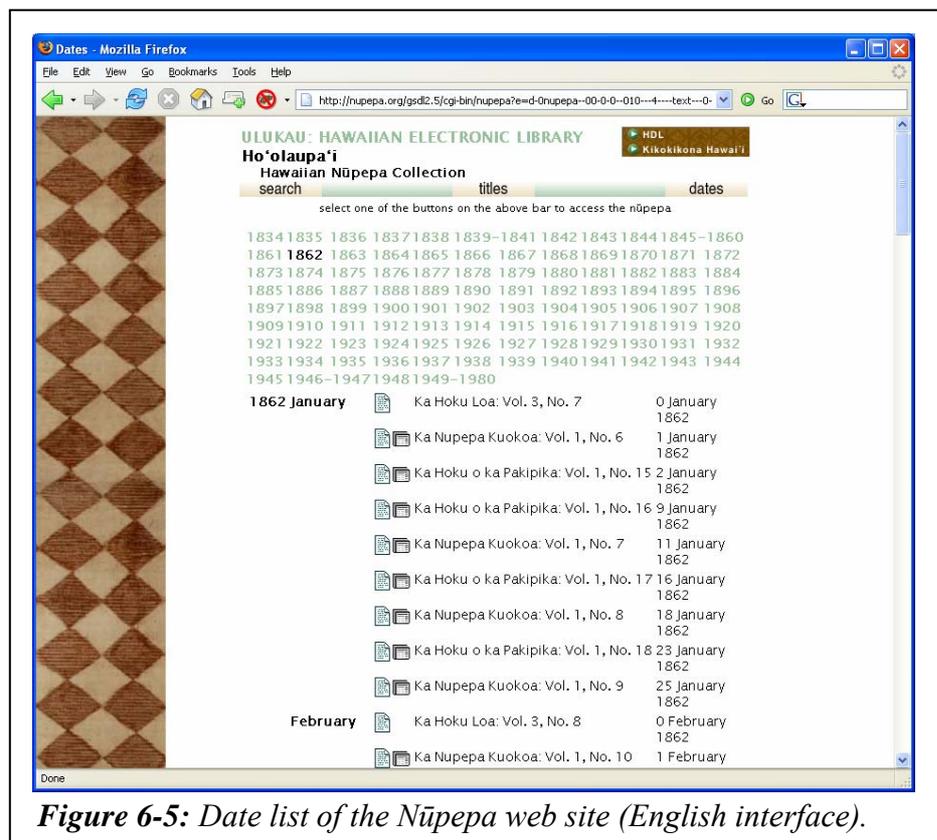


Figure 6-5: Date list of the Nūpepa web site (English interface).

6.1.4 Nūpepa/Niupepa Interface Differences

While the Hawaiian Nūpepa digital library shares some significant similarities with the Māori Niupepa digital library there are some important interface differences. These interface differences affect how Hawaiian Nūpepa clients and Māori Niupepa clients interact with their respective digital libraries and should be kept in mind when usage of these interfaces is compared. The differences are summarised in *Table 6-1*:

The first item listed in *Table 6-1* is the disabling of browser caching. The Ulukau web server was set up to disable browser caching. The HTTP headers that are sent with the web pages have a 'must revalidate' command written in them that instructs the users' browsers not to cache these pages. This feature was not used in the Māori Niupepa web site. It is difficult to determine how effective this command is at disabling browser and network caching and the subsequent affect it has on web log statistics.

Interface Features of Nūpepa and Niupepa Digital Libraries		
	Hawaiian Nūpepa	Māori Niupepa
Disable browser caching	Yes	No
Preferences page	No	Yes
Help page	No	Yes
Interface languages available	2	43
Language switch available on all pages	Yes	No
Full size images	~40,000	35,482
Mid-size images	~40,000	0
Preview image	~40,000	35,482
Extracted text generated from images	~6.4%	98.6%
Detach text	Yes	No
Parallel (indigenous/non-indig.) documents	<1%	44.2%
English language documents	0%	1.9%
Abstracts in English	0%	43.5%
Commentary information	0%	64.7%

Table 6-1: Comparison of Nūpepa and Niupepa interface features as available to clients in 2005.

As shown in *Table 6-1* some features available in the Māori Niupepa digital library were not available in the Hawaiian Nūpepa digital library. The features not available in the Hawaiian Nūpepa digital library were: a preferences page, a help page, a multilingual interface, a high proportion of extracted text documents, a significant number of parallel (indigenous language/non-indigenous language) documents, Abstracts in English and commentary information.

The Hawaiian Nūpepa digital library had some features that were not available in the Māori Niupepa digital library. These extra features included: the disabling of browser caching, the ability to switch the language of the interface (Hawaiian to English and English to Hawaiian) on every interface page, the option to detach (open) the text in another browser window, mid-size images and the ability to view full size images in a pdf format. The pdf format gave the client the ability to zoom and pan on the image, and allowed easier printing of the image.

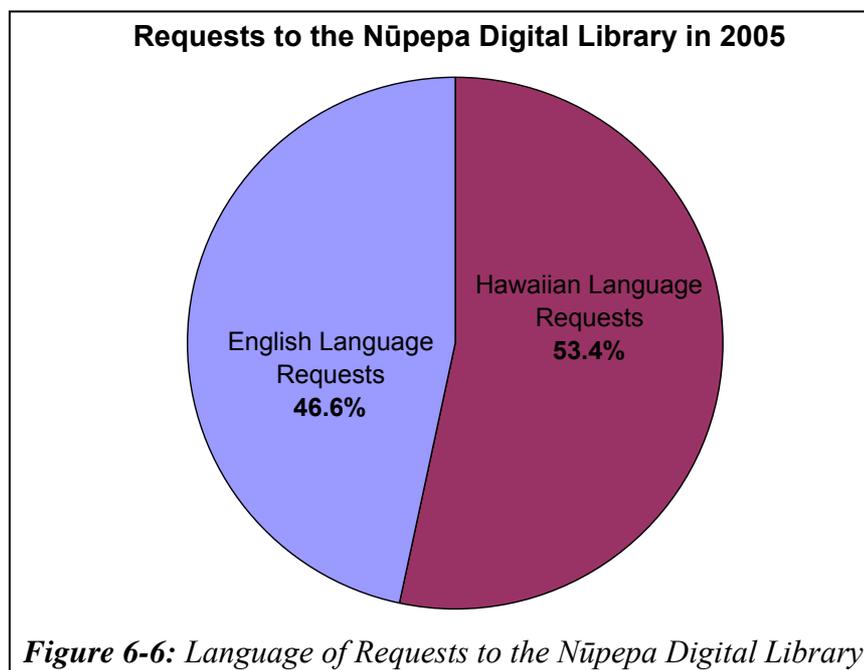
Two further features of the Hawaiian Nūpepa digital library meant it was more likely than the Māori Niupepa digital library to encourage the use of the indigenous language. First, virtually all the content was available only in the Hawaiian language. This meant that the resource was of prime benefit to Hawaiian-speaking clients and of little benefit to clients who could not understand Hawaiian. In contrast, the Māori Niupepa digital library included a significant proportion of English-language material, which benefited clients who only spoke English. The material comprised 44% parallel language (reo Māori/English) Niupepa documents, 2 % newspaper documents published in English, and 44.5% of reo Māori documents that had additional summary abstracts written in English.

The default language setting was the second feature of the Hawaiian Nūpepa digital library that encouraged use of the indigenous language. The Nūpepa default interface language was set to Hawaiian for the whole of 2005. The previous chapter showed that an indigenous language setting not only increases the number of requests recorded in the indigenous language but also increases indigenous language use of other aspects of the interface.

6.2 Was the Hawaiian Language Used?

Transaction Log Analysis (TLA) of the Hawaiian Nūpepa digital library was used to answer whether the Hawaiian language was used by Nūpepa clients.

The Hawaiian Nūpepa web site received a total of 286,383 raw requests in 2005. The first request was logged at 01:17am January 1st 2005 and the final request was logged at 11:59pm December 31st 2005. This raw log of requests was then prepared by removing all redundant CGI arguments and extraneous requests (using the method described in §2.3), leaving a filtered total of 211,215 requests. Of this total, 112,786 were logged as requests with the interface language set to Hawaiian and 98,429 were requests with the interface language set to English. These requests are displayed as percentages in *Figure 6-6*.

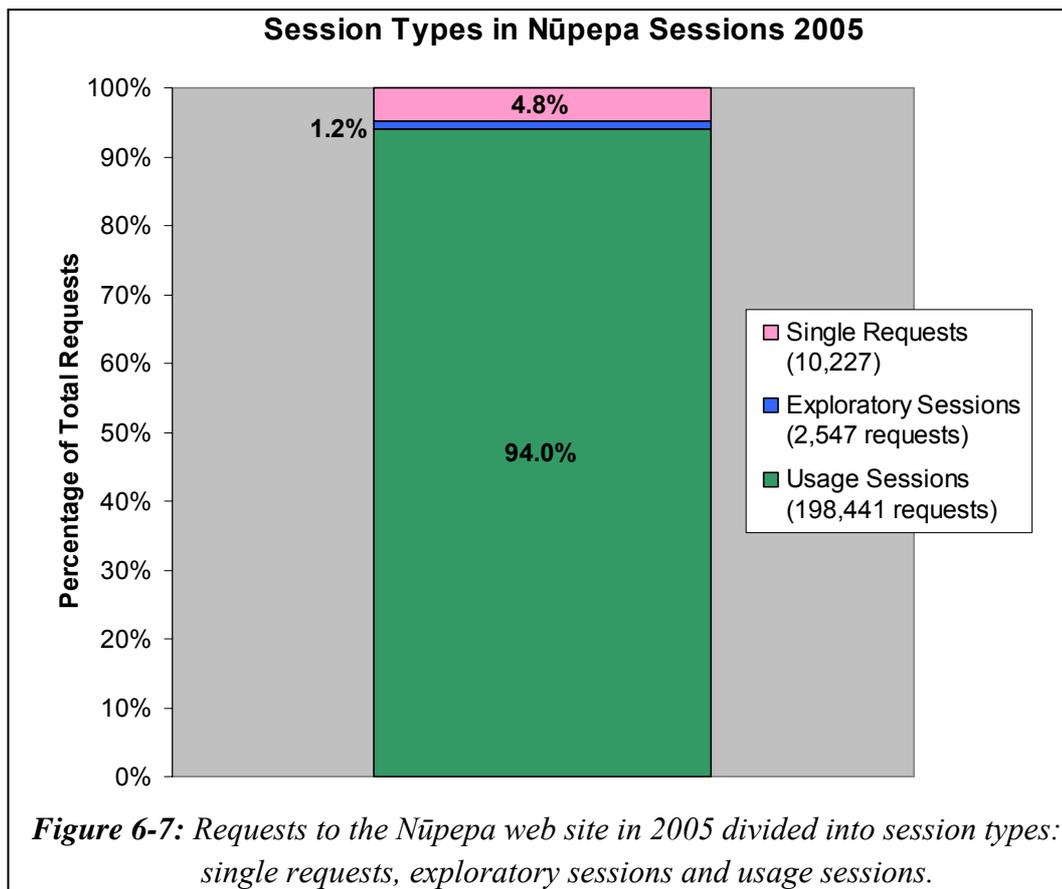


Clearly the Hawaiian language was used by Nūpepa clients; the interface was set to Hawaiian in more than half of the requests made to the Nūpepa digital library.

6.2.1 Was Hawaiian Used in All Three Session Types?

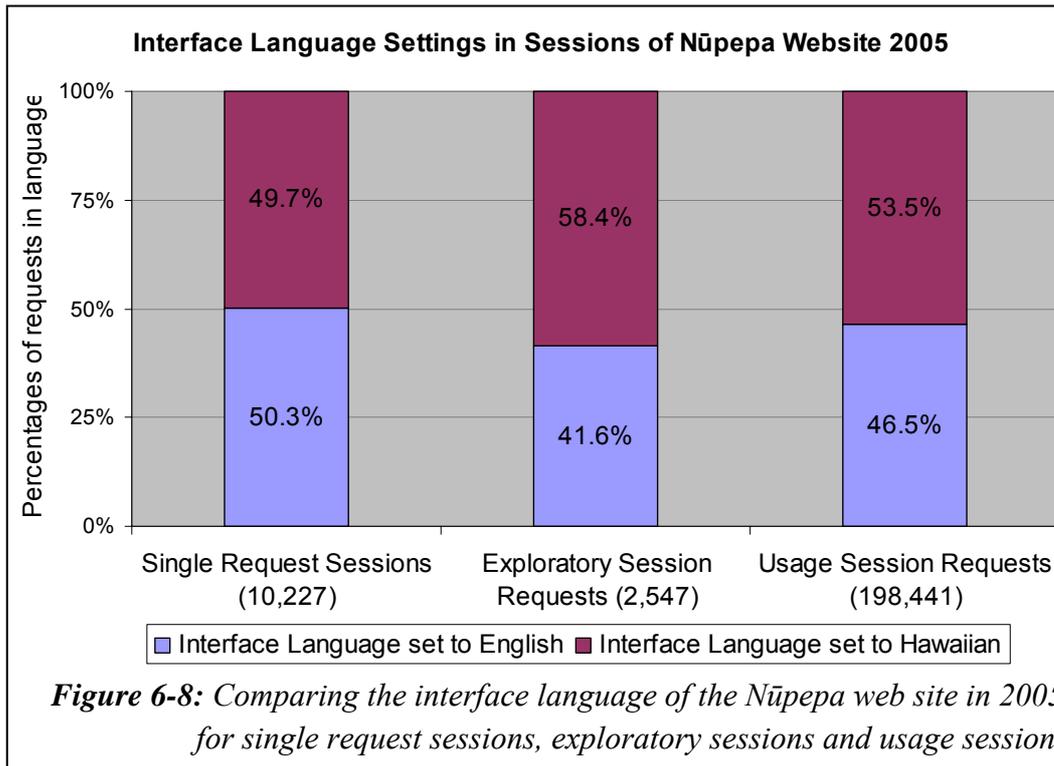
The TLA process collated the requests into three types of sessions. There were 10,227 requests (4.8%) classified as single requests, 2,547 requests (1.2%) classified into 948 exploratory sessions and 198,441 requests (94.0%) classified into 8,754 usage sessions. The percentages of requests per session type are displayed in *Figure 6-7*.

Figure 6-7 shows the majority of the requests (94.0%) to the Nūpepa digital library were classified as usage session requests; requests from clients who were actively engaging with the Nūpepa digital library to retrieve information.



The three session types were further examined to determine which interface language was preferred by Nūpepa clients in 2005. For single requests, 5,085 requests had the interface language set to Hawaiian and 5,142 had the interface

language set to English. For exploratory sessions, 1488 requests had the interface language was set to Hawaiian and 1059 had the interface language was set to English. For usage sessions, 106,213 requests had the interface language set to Hawaiian and 92,228 had the interface language set to English. These figures have been calculated as percentages and are graphed in *Figure 6-8*.



The Hawaiian language was the preferred language of use. *Figure 6-8* clearly shows higher percentages of requests in Hawaiian than in English.

The fact that 53.5% of requests in usage sessions are in Hawaiian is noteworthy as it suggests two important findings. First, not only are Nūpepa clients willing to undertake exploratory and browsing activity in Hawaiian; they are also willing to undertake information retrieval in Hawaiian. Second, it suggests that setting the default interface language to Hawaiian is appropriate for most of the usage activity undertaken in the Nūpepa digital library.

6.2.2 Was the Hawaiian DL Used Bilingually?

Clients could use the Nūpepa digital library in Hawaiian, in English or in both languages, that is, bilingually. A bilingual Nūpepa session was defined the same as a bilingual Niupepa (Māori) session (see §2.4.3). There were 1,090 bilingual sessions in 2005, resulting in 50,359 usgBi requests to the Nūpepa digital library. Bilingual sessions comprised 13,541 (26.9%) requests using Hawaiian and 36,818 (73.1%) requests using English. Bilingual sessions made up 12.5% of the total usage sessions and 25.4% of the total usage requests to the Nūpepa digital library.

The Hawaiian Nūpepa digital library was used bilingually. One in eight usage *sessions* and one in four usage *requests* were categorised as bilingual.

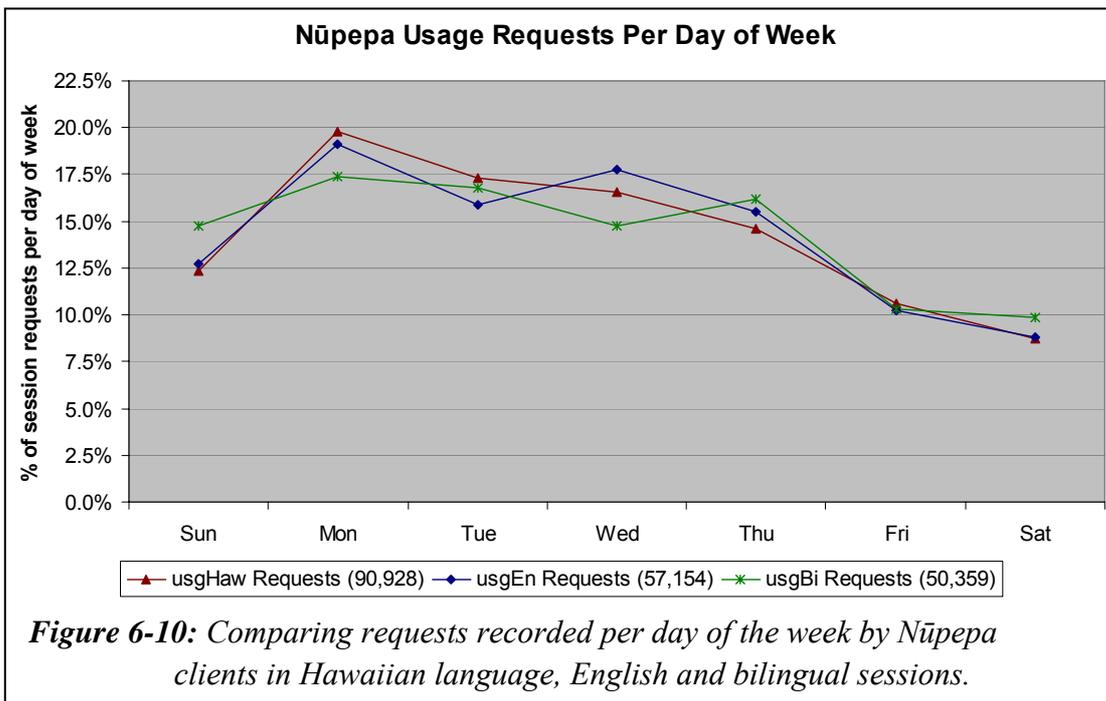
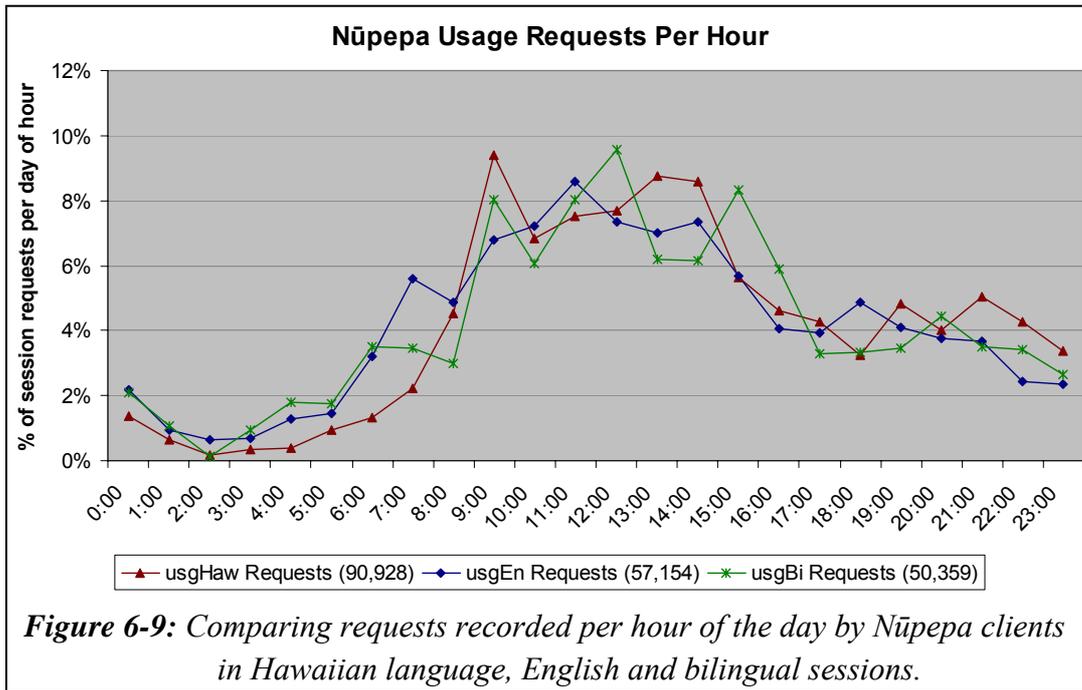
6.2.3 When Was the Hawaiian Language Used?

The times that usage requests were received (time of day, day of week, and month of year) are shown for each language type in *Figures 6-9, 6-10 and 6-11*.ⁱⁱⁱ The sessions have been plotted as percentages of requests per session to take into account the variation in frequency of requests in the different session types.

Figure 6-9 shows the requests as they occurred in a 24 hour time period. The requests occurred at similar times no matter which interface language was being used by the Nūpepa clients. The 9 hours between 8:00am and 5:00pm, the standard work hours, was when most of the activity occurred; 62% of all requests were logged in this time period.

Figure 6-10 shows the requests as they occurred each day. Most of the activity occurred Monday to Thursday, with the least activity occurring on a Saturday.

ⁱⁱⁱ Note: no allowance was made in these graphs for the 6.7% of clients who were making requests from a different time/date zone.



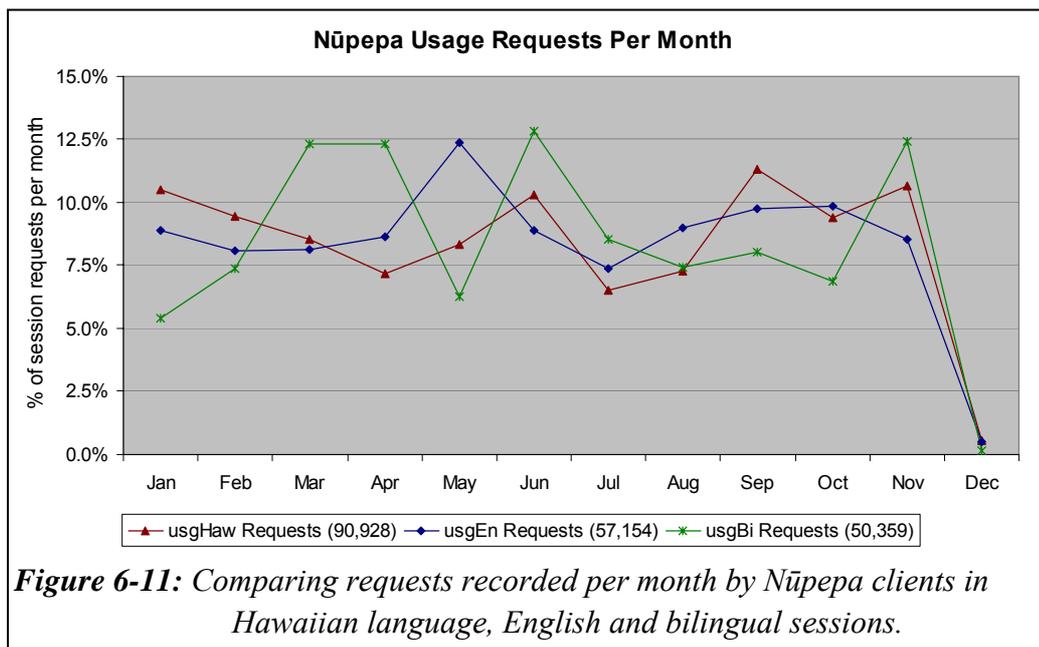


Figure 6-11 shows the requests as they occurred throughout 2005, grouped by month. The possibility of some data being compromised because of a configuration setting of the Ulukau web server, as described in §2.2.4, led to the removal of almost all the December requests from this analysis.

A similar proportion of Hawaiian (usgHaw) and English language usage sessions (usgEn) occurred per month. While the proportion of bilingual usage sessions (usgBi) per month was not as closely aligned, the difference was never more than 6.2%. The Nūpepa interface language was consistently used in Hawaiian, English and bilingually throughout 2005.

6.2.4 Summary—Was the Hawaiian Language Used?

The Hawaiian language was used in more than half of the requests submitted by Nūpepa clients. Approximately half of the Nūpepa usage requests were recorded as usgHaw sessions, one quarter as usgEn sessions and one quarter as usgBi sessions. UsgHaw, usgEn and usgBi clients used the Nūpepa digital library in similar chronological periods; mostly 9:00am to 5:00pm, Monday to Thursday, in all months of the year.

6.3 Exploratory Behaviours

Requests made by clients who explored the Nūpepa digital library but then did not go on to use it represent only 1.2% of the total requests. In total, 2547 requests were defined into 948 Nūpepa exploratory sessions. Nūpepa exploratory sessions were usually short sessions; a high proportion (84.0%) was less than three minutes and (93.7%) consisted of three requests or less. Two requests were the minimum requirement for classification as a session.

Nūpepa exploratory sessions were subdivided into two language categories. There were 1,589 requests in 537 exploratory sessions where the interface language was mostly set to Hawaiian (expHaw). In these sessions the interface language was set to Hawaiian in 72.6% of the requests. There were 958 requests in 411 exploratory sessions where the interface language was mostly set to English (expEn). The interface language was set to English in 64.6% of these requests.

What information can be gained from the TLA process to illustrate how the Nūpepa digital library was explored but then not used by Hawaiian language clients? The Hawaiian Nūpepa interface was simpler than the Māori Niupepa interface as there was no help page or preferences page. This meant that an analysis of which interface pages were requested could not be undertaken. Only two analyses could be made, the length of exploratory sessions and the apparent read times in exploratory sessions.

6.3.1 How Long Were Exploratory Sessions?

Session length is measured in two categories, duration in time (seconds) and length in number of requests. The comparative expHaw and expEn lengths of sessions in seconds are displayed in *Table 6-2*. The comparative expHaw and expEn lengths of sessions in number of requests are displayed in *Table 6-3*.

expHaw			expEn		
Seconds	Sessions	%	Seconds	Sessions	%
Total	537	100%	Total	411	100%
0–1	51	9.5%	0–1	10	2.4%
2–9	49	9.1%	2–9	146	35.5%
10–29	205	38.2%	10–29	133	32.4%
30–179	163	30.4%	30–179	80	19.5%
180+	69	12.8%	180+	42	10.2%

Table 6-2: Comparisons of lengths (in seconds) of expHaw and expEn sessions.

expHaw			expEn		
Requests	Sessions	%	Requests	Sessions	%
Total	537	100%	Total	411	100%
2	177	33.0%	2	343	83.5%
3	270	50.3%	3	41	10.0%
4	51	9.5%	4	15	3.6%
5+	39	7.3%	5+	12	2.9%

Table 6-3: Comparisons of number of requests in expHaw and expEn sessions.

ExpHaw sessions were short; 56.8% were less than 30 seconds and 87.2% were less than 3 minutes. In terms of number of requests 83.3% consisted of just two or three requests.

Both tables show that expHaw sessions were slightly longer than expEn sessions. *Table 6-2* indicates that expHaw sessions had a higher proportion of 30–179 second sessions (30.4% compared with 19.5%). The average duration of expHaw sessions was 1.63 minutes while the average duration of expEn sessions was 1.56 minutes. *Table 6-3* indicates that expHaw sessions had a much higher proportion of sessions involving three requests (50.3% compared with 10.0%). The average number of requests in expHaw sessions was 3.0 requests compared with 2.3

requests in expEn sessions. While these differences may not appear large, with p values of $3.9 * 10^{-23}$ and $2.0 * 10^{-52}$ respectively, the differences were statistically significant.

While both expHaw and expEn sessions were short, the expEn sessions were shorter than the expHaw sessions. The difference between the length of expHaw sessions and the length expEn sessions was statistically significant.

6.3.2 How Long to Read Interface Pages?

The lengths of times between requests for pages to the Nūpepa web site were calculated and recorded as apparent page read times. These read times were grouped into time intervals of, 1–2 seconds, 3–9 seconds, 10–19 seconds, 20–39 seconds, 40–79 seconds and 80+ seconds. These times are shown in *Table 6-4*.

The average apparent read time of interface pages by expHaw clients was 61.4 seconds. The average apparent read time of interface pages by expEn clients was 83.8 seconds. Chi square analysis on the read times listed in *Table 6-4* returned a p value of 0.22. This p value indicates that there was no statistically significant difference in the expHaw and expEn apparent read times of Nūpepa interface pages.

Seconds	expHaw	expEn
Total Apparent Reads	1052	547
1–2	103	37
3–9	362	195
10–19	257	132
20–39	145	79
40–79	56	41
80+	129	63

Table 6-4: expHaw and expEn apparent read times of Nūpepa interface pages in exploratory sessions.

6.3.3 Summary—Exploratory Behaviours

Exploratory sessions in Hawaiian, like exploratory sessions in te reo Māori, were short. ExpHaw sessions mostly consisted of two or three requests and half of the sessions were completed within 30 seconds. Almost half of the apparent read times of interface pages by expHaw clients were less than 10 seconds.

While exploratory activity by Hawaiian language clients was short, it wasn't as short as exploratory activity by English language clients. ExpEn clients undertook shorter exploratory sessions, both in duration and number of requests, than expHaw clients. The apparent read times of interface pages by expHaw clients and expEn clients were statistically similar.

Exploratory clients only request pages from the interface; no documents are requested and no queries are submitted. Aside from session lengths and apparent read times, it is difficult to derive information about how the Hawaiian language was used by exploratory clients of the Nūpepa digital library. Exploratory session requests comprised just a small (1.2%) portion of the activity that occurred in the Nūpepa digital library in 2005.

The next three sections analyse how the Hawaiian language was used by clients to actively engage with the Nūpepa digital library to retrieve information. Usage of the Nūpepa digital library by usgHaw clients is analysed in three parts; interface behaviours, document access behaviours and query behaviours. The first part, an analysis of how usgHaw clients used the Nūpepa interface, begins in the following section.

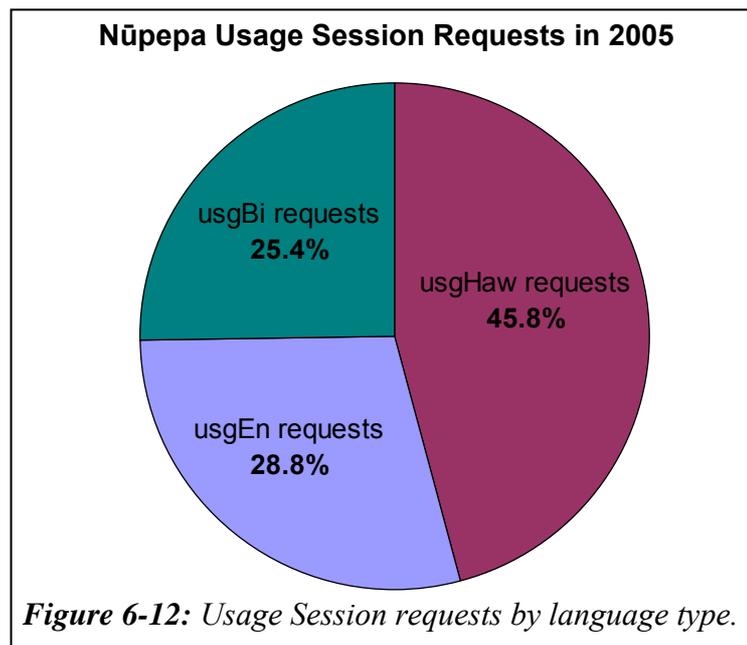
6.4 Nūpepa Interface Behaviours

This section analyses how the Nūpepa interface was used by Hawaiian language clients in usage sessions. A usage session is a session where the client actively engages with the Nūpepa interface to access Nūpepa documents. Data about usage session activity is the most important session data to analyse as it represents the majority (94%) of activity and specifically targets the key target group, clients who actively use the Nūpepa digital library.

There were 198,441 requests recorded in 8,754 usage sessions in the Hawaiian Nūpepa digital library in 2005. Of the requests recorded, 106,213 (53.5%) used the interface in Hawaiian and 92,222 (46.5%) used the interface in English. On average, the usage sessions undertaken by Nūpepa clients were 10 times longer in duration and eight times longer in number of requests than the exploratory sessions.

The Nūpepa usage session requests were then subdivided into language type (as described in §2.4.3). There were 90,928 requests in 4,182 usage sessions when the language of the interface was mostly set to Hawaiian (usgHaw) and 57,154 requests in 3,482 usage sessions when the language of the interface was mostly set to English (usgEn). There were 50,359 requests in 1,090 usage sessions when the language of the interface was switched between Hawaiian and English; these sessions were subsequently defined as bilingual sessions (usgBi). An analysis of the usgHaw requests shows the interface language set to Hawaiian in 98.6% of the requests. An analysis of the usgEn requests shows the interface language set to English in 94.7% of the requests. An analysis of the usgBi requests show the interface set to English in 73.1% of the requests and Hawaiian in 26.9% of the requests.

The usage session requests for each language type were calculated as percentages and are displayed in *Figure 6-12*.



Statistics generated from TLA were used to examine six interface behaviours exhibited by Nūpepa clients who used the Hawaiian language:

- the length of usgHaw sessions
- the apparent read times of interface pages in usgHaw sessions
- the number of return usgHaw sessions
- the origin of usgHaw sessions
- the frequency of interface language switching
- the position of interface language switching in usgHaw sessions.

These six interface behaviours comprise the next six sub-sections.

6.4.1 How Long Were Usage Sessions?

Session length is defined by two measurements; duration in time (seconds and minutes) and length in number of requests. Statistics generated from TLA on the Nūpepa digital library are listed to show the length of usgHaw sessions in 2005. These statistics are listed alongside length statistics on usgEn and usgBi sessions.

How Long in Time?

The durations of usgHaw sessions were grouped into the following time intervals; 0–10 seconds, 11–60 seconds and 1–3 minutes, 3–15 minutes, 15–30 minutes, 30+ minutes. The numbers of sessions that fitted into these categories were calculated as percentages for comparison purposes and graphed in *Figure 6-13*.

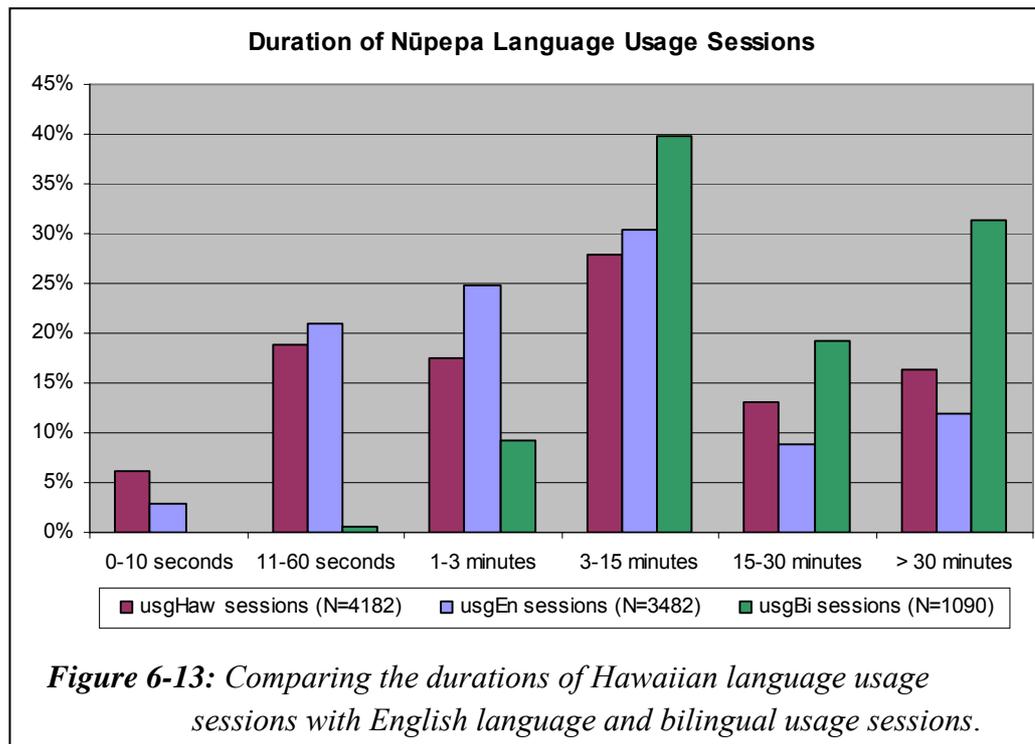


Figure 6-13 shows that usgHaw and usgEn sessions appeared to have similar session durations, while the usgBi sessions were longer. However, when chi square analysis was undertaken on the session lengths of usgHaw and usgEn sessions it generated a p value of $5.0 * 10^{-27}$. This indicates that the difference is statistically significant. UsgHaw sessions had a higher proportion of longer sessions (in the 15–30 minutes and >30 minute category) while usgEn sessions had a higher proportion of shorter sessions (in the session categories 15 minutes or less).

The average usgHaw session duration was 16.3 minutes, the average usgEn session was 12.4 minutes. The usgHaw sessions were, on average, a third longer

than the usgEn sessions but only half as long as the usgBi sessions. By definition a usgBi session must involve at least ten requests so it is expected that these sessions will be long.

How Long in Number of Requests?

Also analysed was the length of usage sessions in terms of number of requests by clients who used the Nūpepa digital library in Hawaiian. The numbers of requests per usgHaw session were grouped into the following divisions; 2 requests, 3–4 requests, 5–8 requests, 9–20 requests, 21–40 requests and 41+ requests. The numbers of sessions per request count category were calculated as percentages and are displayed in *Figure 6-14*.

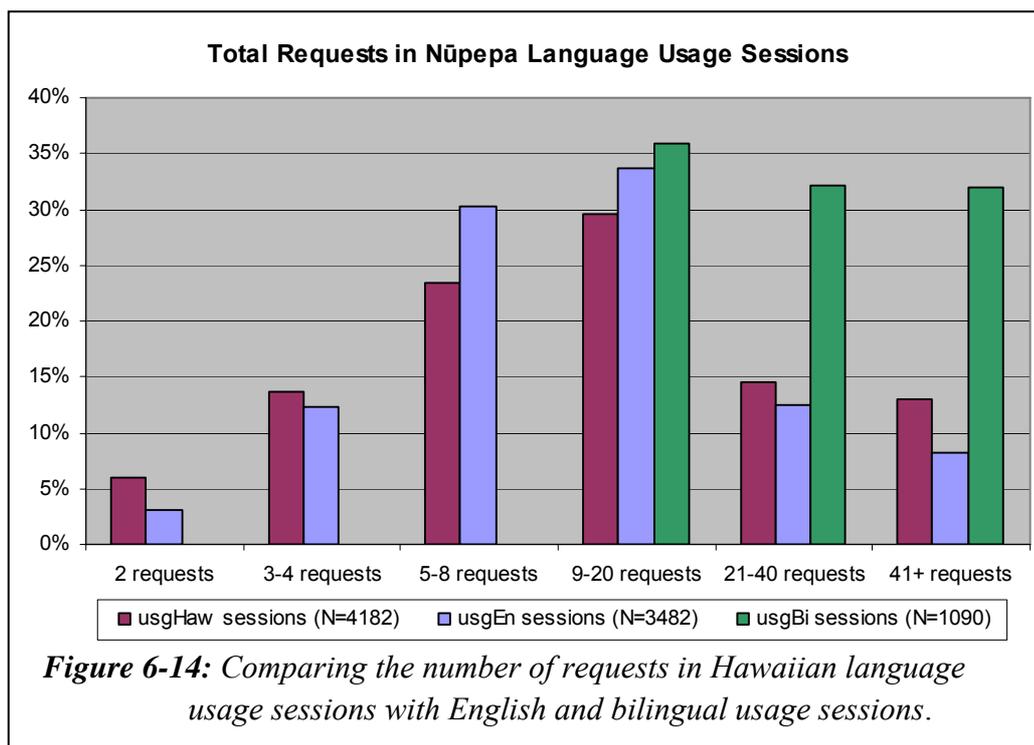


Figure 6-14 shows that there were more short sessions (2 requests and 3–4 requests) and more long sessions (21–40 requests and 41+ requests) in Hawaiian than there were in English. There were proportionally more mid-range (5–8 requests and 9–20 requests) in English than in Hawaiian. UsgHaw sessions had an

average of 21.7 requests and a median of 10 requests per session. The usgEn sessions had an average of 16.4 requests and a median of 9 requests per session.

The statistics for usgBi sessions indicated that these sessions had, on average, over twice as many requests as usgHaw sessions. Again, this high number may simply have occurred because of the criteria for determining a usgBi session; these sessions must be at least 10 requests long.

6.4.2 How Long to Read Interface Pages?

Apparent interface page read times are used to determine how long usgHaw clients took to read pages of the Nūpepa interface. The apparent read times in usgHaw, usgEn, usgBi and in total usage sessions were calculated and descriptive statistics are shown in *Table 6-5*.

Apparent Page Read Times of Interface Pages in Usage Sessions				
	usgHaw	usgEn	usgBi	Total
Total Apparent Reads	4948	5302	3298	13548
Average (seconds)	37.9	37.0	44.9	39.3
Standard Deviation	183.8	179.6	188.1	183.3
Mode (seconds)	3	4	4	3
Shortest session (seconds)	0	0	0	0
Lower Quartile (seconds)	3	5	4	4
Median (seconds)	6	9	9	8
Upper Quartile (seconds)	15	20	21	19
Longest session (seconds)	3492	3489	3412	3492

Table 6-5: Summary statistics of apparent read times of interface pages in usgHaw, usgEn and usgBi sessions.

UsgHaw clients took, on average, 37.9 seconds to read interface pages of the Nūpepa digital library. This average time was over 1/3 quicker than the time it

took expHaw clients to read interface pages in exploratory sessions (61.4 seconds).

The average apparent read time undertaken by usgHaw clients appears similar to the average apparent read time undertaken by usgEn clients. The high standard deviations listed in *Table 6-5* indicate that there is a wide variation in the usgHaw and usgEn apparent read times. The usgBi sessions took the longest time between page requests.

6.4.3 How Many Clients Returned?

Statistics generated from TLA can indicate how often clients returned to the Nūpepa digital library. The statistics are generated under the following assumptions: clients always had cookies enabled, the cookies were not routinely deleted, clients always logged on using the same account on the same computer, and because browser caching was disabled by the Ulukau web server, there was minimal to nil network and local browser caching. While these assumptions were not valid for some users in some sessions, the figures can still give comparative indications on the number of return clients.

The number of clients who undertook return sessions and the number of requests they submitted to the Nūpepa web site were collated. From a total of 6,839 identifiable clients there were 555 (8.1%) clients who undertook return sessions. These clients made 60,042 (30.3%) requests for pages and documents from the Nūpepa web site. While only a small number of clients were identified as return clients, they were responsible for a considerable percentage of the requests.

The return analysis was also undertaken for each client language type. There were 378 (12.9%) usgHaw clients who undertook return sessions, requesting 98,545 (63.6%) pages and documents. There were 270 (8.7%) usgEn clients who undertook return sessions, requesting 15,353 (23.7%) pages and documents. There were also 137 (13.0%) usgBi clients who undertook return sessions requesting 8,617 (16.7%) pages and documents in usgBi sessions.

These figures indicate some important points. First, usgHaw and usgBi clients were the clients most likely to return to the Nūpepa web site, returning one and a half times more often than usgEn users. Second, when usgHaw clients returned to the Nūpepa web site they undertook significantly more requests than usgEn and usgBi clients. There were 21.8 requests submitted per non-returning usgHaw client and 329.3 requests submitted per returning usgHaw client. By comparison, there were 17.5 requests submitted per non-returning usgEn client and 78.2 requests per returning usgEn client. Also there were 46.2 requests per non-returning usgBi client and 153.6 requests per returning usgBi client.

Some clients undertook sessions in different language categories. There were:

- 24 clients recorded in all three session types, usgHaw, usgEn and usgBi
- 79 clients recorded in both usgHaw and usgEn session types
- 42 clients recorded in both usgHaw and usgBi session types
- 61 clients recorded in both usgEn and usgBi session types.

While it was noted that some clients undertook sessions in different language categories, it was not considered significant. For example, the 24 clients who undertook sessions categorised as usgHaw, usgEn and usgBi sessions represented only 0.4% of the total number of unique clients.

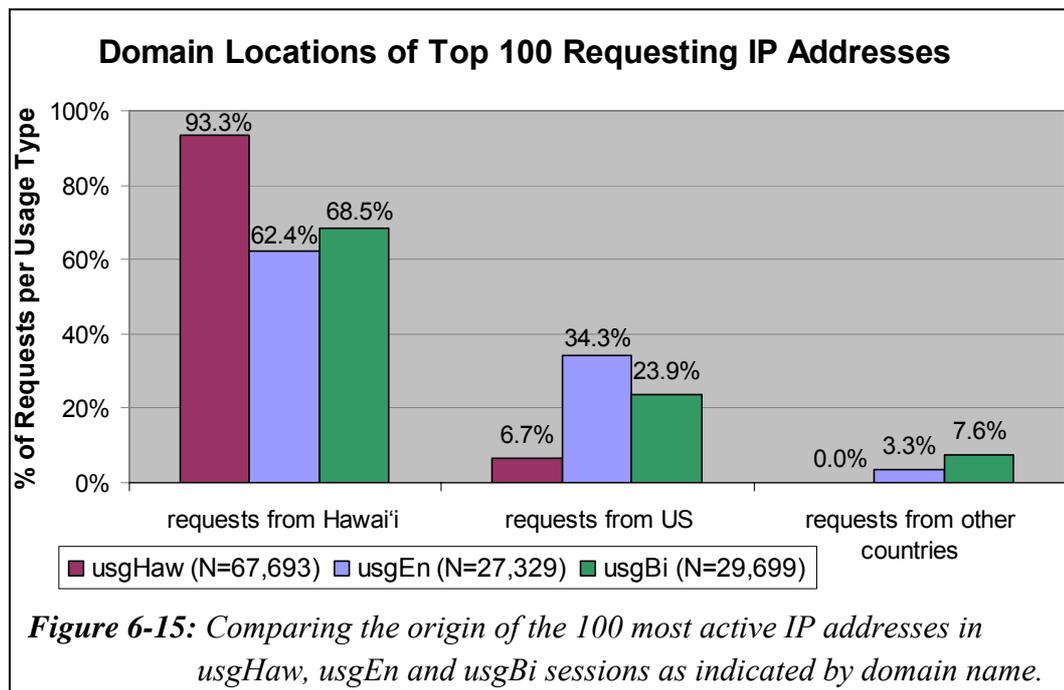
6.4.4 Where Did Usage Sessions Originate?

The TLA process can give information about the origin of requests, as described in §3.4.4. A process called reverse DNS lookup can be used to return information about the computer/network device including domain name and location.

Reverse DNS lookup was undertaken on the 100 IP Addresses which made the most requests in each usage sessions type. The 100 most active usgHaw IP Addresses submitted 67,693 requests, the 100 most active usgEn IP Addresses submitted 17,329 requests, and the 100 most active usgBi IP Addresses submitted

29,699 requests. The 100 usgHaw IP Addresses submitted more than twice as many requests as the usgEn or usgBi IP Addresses.

Results from Reverse DNS lookup indicated three general origins; those IP Addresses that appeared to originate from Hawai‘i, those from the United States and those from other parts of the world. The percentages of requests by location for each usage session type are displayed in *Figure 6-15*.



Most of the requests to the Hawaiian Nūpepa digital library originated from Hawai‘i. Hawaiian usage sessions, at 93.3%, had the highest rate of requests originating from Hawai‘i, suggesting that the resource was being used extensively by local people. Given that over 90% of Hawaiian speakers are of Hawaiian ancestry (Wilson, 2006), these figures suggest solid use of this web site by the indigenous Hawaiian people.

No overseas country made it into the top 100 IP requests in Hawaiian usage sessions. For the English usage sessions there were two IP addresses requesting from Aotearoa (a total of 775 requests) and one from the Netherlands (133 requests). The bilingual usage sessions showed the most involvement from

overseas users; there were two IP addresses requesting from Canada (a total of 295 requests), two IP addresses from Aotearoa (a total of 287 requests), one from Germany (897 requests), one from Holland (430 requests) and one from Japan (360 requests). These figures suggest that users from overseas countries are willing to switch the interface language of the Nūpepa web site between the Hawaiian and English languages.

The question posed in this sub-section was, where did requests by Hawaiian language clients originate? The answer is, they mostly originated in Hawai‘i.

6.4.5 How Often Was the Interface Language Switched?

As the default language of the Nūpepa interface was set to Hawaiian it could be expected that only a small number of the usgHaw sessions would involve a language switch. Only 340 (8.1%) of the total 4,182 Hawaiian usage sessions switched the language of the interface. UsgHaw sessions represented 45.8% of the total usage sessions yet were only responsible for 4.8% of the sessions that undertook interface language switching.

Perhaps somewhat surprising was that only 53.4% of the usgEn sessions switched the interface language. This figure suggests that 46.6% of the usgEn sessions had the Nūpepa web site bookmarked in English.

Slightly more of the language interface switches were a switch from the Hawaiian language to English. The Hawaiian-English interface language switch was evident in 56.5% of the usgHaw session switches, 64.6 % of the usgEn session switches and 52.2% of the usgBi session switches.

6.4.6 When Was the Interface Language Switched?

To determine where the interface language switches were occurring in usgHaw sessions, the sessions were divided into 10 equal portions and the number of switches in each portion was counted. *Figure 6-16* shows the percentages of Hawaiian to English switches for usgHaw sessions and *Figure 6-17* shows the percentages of English to Hawaiian switches for usgHaw sessions. *Figures 6-16* and *6-17* also show the percentages of when switching occurred in usgEn and usgBi sessions.

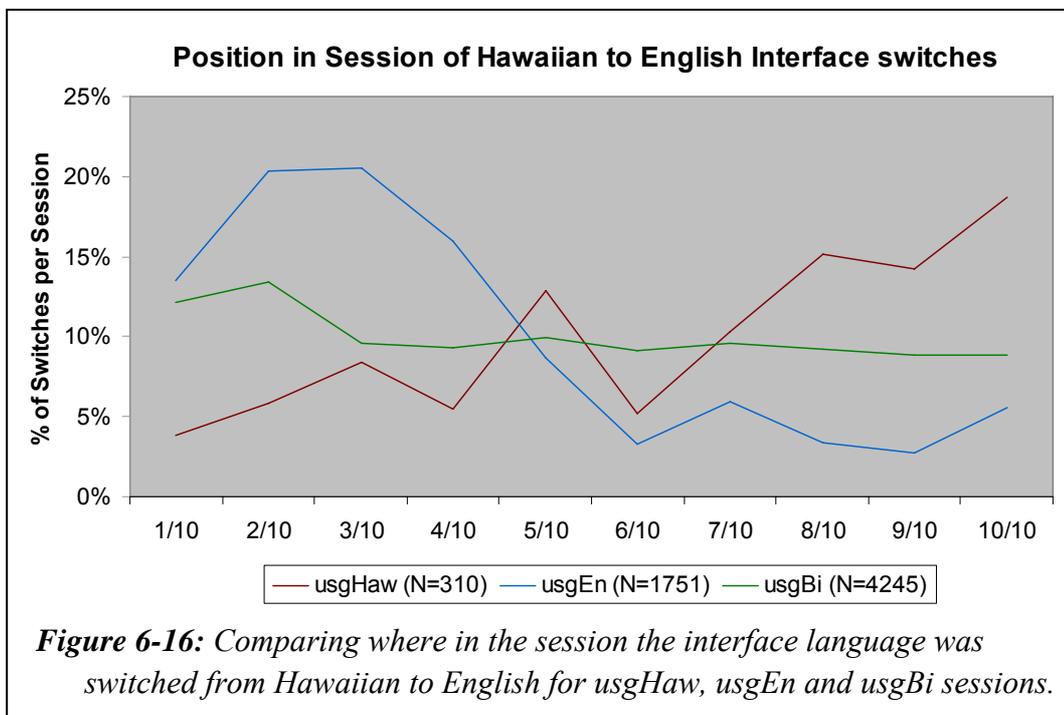
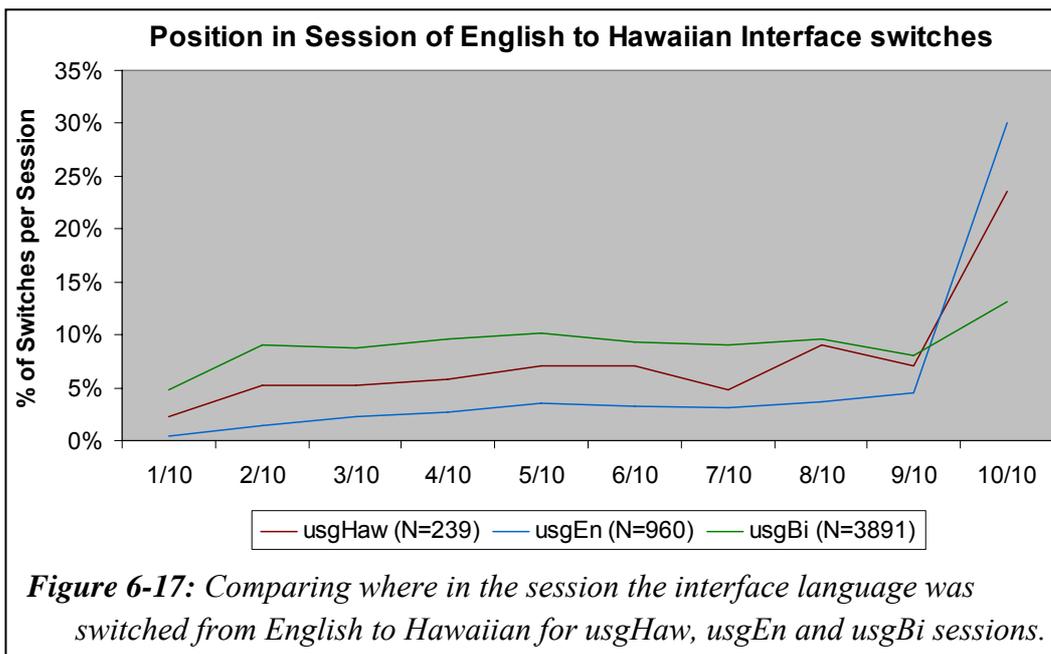


Figure 6-16 shows that when usgHaw clients did switch the language of the interface the switch was most likely to occur towards the end of the session. In contrast, usgEn clients were most likely to switch the interface language from Hawaiian to English at the start of a session, while usgBi clients switched the language of the interface equally throughout the session.

Figure 6-17 shows the interface language switch from English to Hawaiian in usgHaw sessions mostly occurred towards the end of the session. However, it did not occur very often. There were only 239 English to Hawaiian switches in usgHaw sessions, compared with 960 in usgEn sessions and 3,891 in usgBi sessions.



The interface language switch from English to Hawaiian in usgEn sessions also mostly occurred towards the end of the session. However, the interface switch in usgBi sessions occurred at a reasonably constant rate across the whole usgBi session.

6.4.7 Summary—Interface Behaviours

This section analysed how the Nūpepa interface was used by Hawaiian language clients in usage sessions. Six aspects of using the Nūpepa interface in Hawaiian were examined by analysing statistics drawn from TLA.

The statistics revealed that, on average, usgHaw sessions lasted for 16 minutes and consisted of 22 requests. The average apparent read time of an interface page by usgHaw clients was 40 seconds. Almost 13% of the usgHaw sessions were identified as return sessions and these return sessions were responsible for almost 64% of the total usgHaw requests. Most (93%) of the requests submitted by usgHaw clients originated from Hawai‘i. The language of the interface was not switched often by usgHaw clients; when switched, the switch usually occurred towards the end of the session.

This brief summary concludes the examination of how the Nūpepa interface was used in the Hawaiian language. The next stage is to examine how documents were accessed. This analysis begins in the following section.

6.5 Document Access Behaviours

This section analyses how Nūpepa documents were accessed by Hawaiian language clients in usage sessions. By definition (see §2.4.2) if a client requested a document in a session, then that session was regarded as a usage session. Documents available from the Nūpepa digital library consisted of historic Hawaiian language newspaper pages that were served to the client either as a preview image, a mid-size image or a full size image. A small percentage of documents (6.4%) were also available in an extracted text format (see *Figure 6-3*).

TLA was used to provide statistics for seven behaviours exhibited by Hawaiian language clients who accessed documents from the Nūpepa digital library:

- the methods used to access documents in usgHaw sessions
- the document formats favoured in usgHaw sessions
- the Nūpepa periodicals most requested by usgHaw clients
- the publication years most requested by usgHaw clients
- the propensity for documents with associated text files

- the use of the Detach option by usgHaw clients
- the apparent read times of documents by usgHaw clients.

These seven document access behaviours comprise the next seven sub-sections.

6.5.1 How Were Documents Accessed?

As mentioned in §6.1.3 there were three methods by which Nūpepa clients could access documents; they could be requested following a query activity, a browse by series, or a browse by date. When a client requested a document by one of these methods, say a query, and then proceeded to browse documents in the collection, the subsequently browsed documents were recorded as being generated from the query activity. The methods used by usgHaw clients to access documents are displayed in *Table 6-6*, along with the methods used by usgEn and usgBi clients.

	usgHaw		usgEn		usgBi	
Total	48,455	100.0%	20,450	100.0%	24,205	100.0%
From query	13,018	26.9%	12,114	59.2%	12,591	52.0%
From series	31,030	64.0%	4,542	22.2%	5,682	23.5%
From date	4,407	9.1%	3,794	18.6%	5,931	24.5%

Table 6-6: Counts and percentages comparing how documents were accessed by usgHaw, usgEn and usgBi clients.

Clients who used the Hawaiian language (usgHaw clients) preferred to use the series list to access documents. This could indicate that these clients had prior knowledge about specific documents and used the series list to access them directly. Hawaiian language clients (like te reo Māori clients) showed the least interest in accessing documents using a list by date. The query facility might have been used more by usgHaw clients if a higher number of Nūpepa documents had had extracted text files.

UsgEn and usgBi clients preferred to use the query facility to access Nūpepa documents even though only 6.4% of the documents were searchable.

6.5.2 Which Document Formats Were Favoured?

The Nūpepa documents stored in the Nūpepa digital library were made available to the clients in one of four formats; a preview image, a mid-sized image, a full size image in a pdf format and, for a small percentage of documents, an extracted text format. *Table 6-7* lists the total numbers and percentages of these four formats requested by clients.

Formats of Requested Documents						
	usgHaw		usgEn		usgBi	
Total Documents	48,455	100.0%	20,450	100.0%	24,205	100.0%
Preview Image	16,101	33.7%	4,419	22.5%	9,048	38.2%
Mid-sized image	13,148	27.5%	2,082	10.6%	1,870	7.9%
Full Image	2,025	4.2%	535	2.7%	2,518	10.6%
Extracted Text	16,517	34.6%	12,618	64.2%	10,232	43.2%

Table 6-7: Document formats requested by usgHaw, usgEn and usgBi clients.

Preview images and extracted text documents were the formats most requested by usgHaw clients. Approximately 1/3 of the documents requested by usgHaw clients were preview images, 1/3 were extracted text documents and the remaining 1/3 comprised the mid-size and full size images. Considering that only 6% of the documents in the Hawaiian Nūpepa collection had extracted text formats available it is surprising that this format was requested so much.

The extracted text format was the format prominently requested by usgEn clients. This is, perhaps, a result of usgEn clients' preference for using the query facility to access documents. UsgBi clients often exhibited format preferences that were between the extremes of the usgHaw and usgEn clients; usgBi clients preferred

the extracted text format (but not as much as usgEn clients) and also showed the highest preference for preview images.

6.5.3 Which Nūpepa Periodicals Were Favoured?

Identifying which documents of the Nūpepa collection were requested wasn't a straightforward process. The Greenstone software on which the Nūpepa collection was built, had been configured to allocate a unique identifier to each document and subsequently use this identifier when servicing requests for the document. Each unique identifier was created by hashing^{iv} the metadata of the document with any extracted text which had been generated for the document. The unique identifiers were created each time the Nūpepa collection was rebuilt, which happened when new extracted text files or new image files were added to the collection. When the Nūpepa server recorded requests for documents in the Nūpepa web log it recorded the unique identifier of each document. To determine exactly which documents had been requested, the hashing process had to be reversed.

The Nūpepa web log for 2005 was analysed and the unique identifiers extracted. The hashing of unique identifiers for documents was reversed to determine exactly which documents the identifiers were referring to. Unfortunately, changes had occurred in some document details after they were recorded in the 2005 logs. These changes meant that new identifiers were created and the old identifiers, as recorded in the 2005 logs, no longer mapped directly to a document. Of the 93,110 documents requested by clients in 2005, 21,543 (23.1%) did not map directly to a document and therefore could not be used in the document analysis. The remaining 71,567 documents were further analysed to determine exactly which periodicals were accessed by clients who used the Nūpepa digital library in Hawaiian.

^{iv} Hashing is a process where a string of characters has a mathematical process performed on it to produce a unique identifier.

Although 46 periodical titles are currently available on the Nūpepa web site, in 2005 there were fewer than this. No specific records were kept but the web site developers suggest that perhaps there were less than ten periodicals available in 2005. The list of periodicals requested by Nūpepa usage clients, as recorded by almost 77% of the 2005 Nūpepa web log, is displayed in *Table 6-8*.

The most popular periodical was *Ka Nupepa Kuokoa*. With 51,328 requests, this periodical was requested in 71.7% of identified requests, which is perhaps not surprising as the periodical had over 90% of all available pages. The second most requested periodical was *Ka Lama o Hawai‘i* which was requested 8,361 times and accounts for 11.7% of the identified requests. The third and fourth most requested periodicals were *Ka Hoko o Ka Pakipika* at 4,104 requests (5.7%) and *Ka Nonanona* at 3,223 requests (4.5%). No other single periodical was responsible for more than 2% of the identified requests.

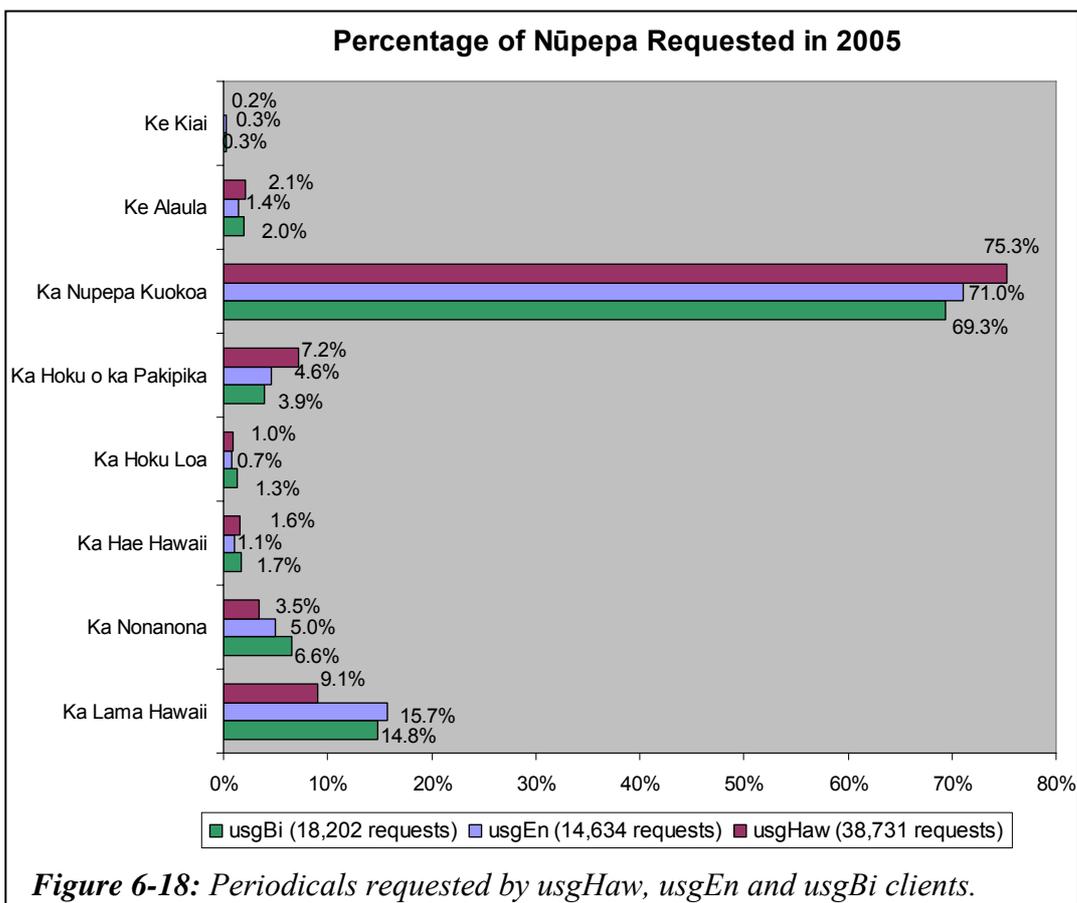
Nūpepa Periodicals Requested in 2005		
Catalogue Number and Name	Requests	Pages Available ^v
001 Ka Lama Hawaii	8,361	109
004 Ka Nonanona	3,223	533
008 Ka Hae Hawaii	1,089	36
009 Ka Hoku Loa	713	104
020 Ka Hoku o ka Pakipika	4,104	325
021 Ka Nupepa Kuokoa	51,328	18,153
024 Ke Alaula	1,370	337
080 Ke Kiai	166	16

Table 6-8: Number of requests per Nūpepa periodical.

^v These were the number of pages available in April 2007. While no accurate records were kept it is suspected that a similar number were available in 2005.

The eight periodicals that were requested in 2005 were then further analysed to see which ones were favoured by usgHaw clients. The percentages of requests per periodical were graphed for usgHaw, usgEn and usgBi clients. These percentages are displayed in *Figure 6-18*.

The periodical *Ka Nupepa Kuokoa* was requested in 75% of requests by clients who used the Nūpepa digital library in Hawaiian. This periodical was slightly (4.3%) more popular with usgHaw clients than it was with usgEn and usgBi clients. The second most favoured periodical by usgHaw clients was *Ka Lama Hawaii* followed by *Ka Hoku o ka Pakipika*.



UsgEn and usgBi clients appeared to have similar preferences for accessing certain periodicals. *Ka Nupepa Kuokoa* was favoured by both client types, followed by *Ka Lama Hawaii*. However, a chi square analysis on the number of

usgEn and usgBi requests for the different periodicals returns a p value of $5.9 * 10^{-21}$. This p value indicates that periodicals requested by usgEn clients were different from those requested by usgBi clients and that the difference was statistically significant.

6.5.4 Which Publication Years Were Favoured?

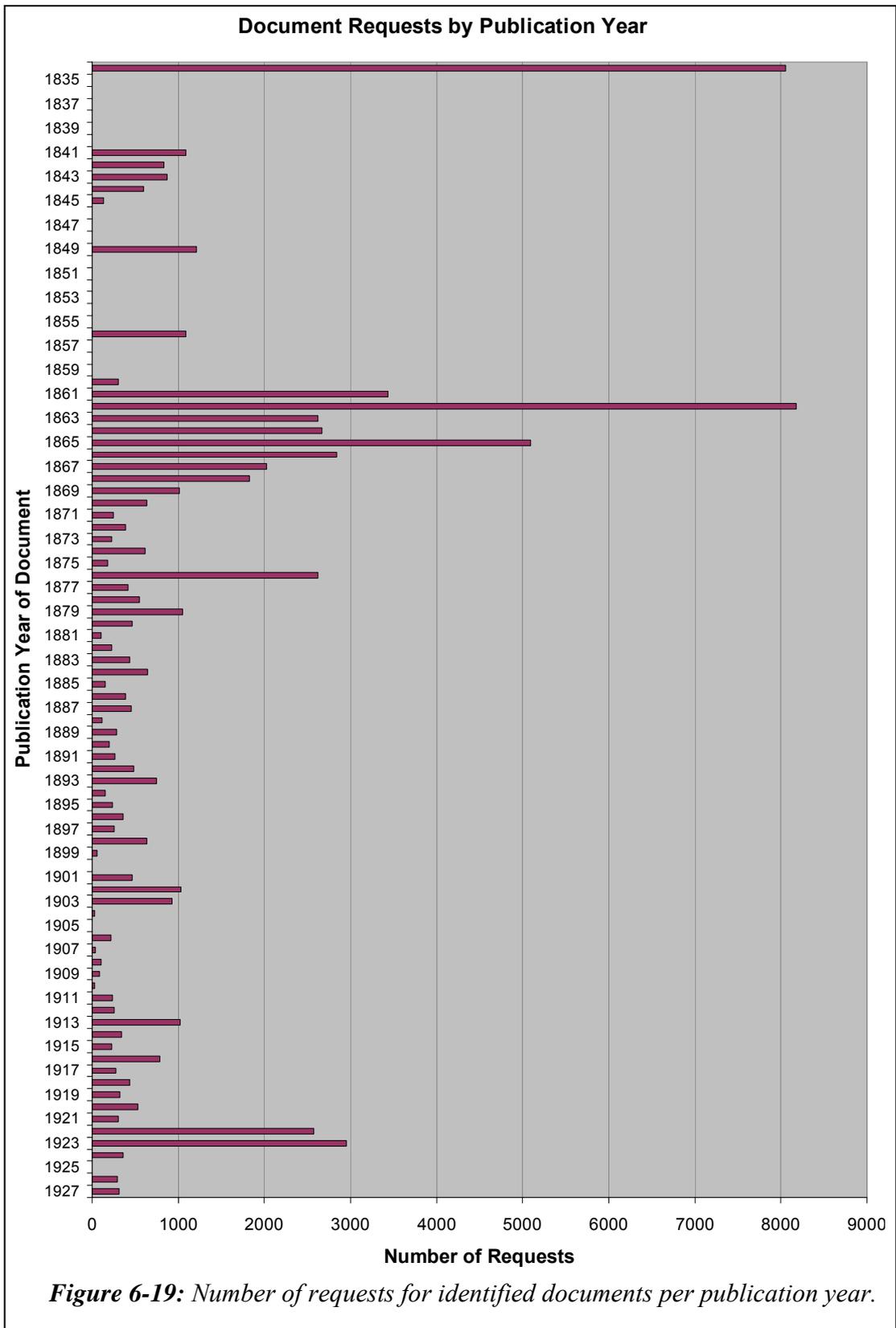
As mentioned in the previous sub-section, 71,567 requested Nūpepa documents could be identified. These identified documents represent 76.9% of the total documents requested by Nūpepa clients. These identified documents were further analysed to determine what year of publication was favoured by Nūpepa clients.

The periodicals in the Nūpepa collection were published over a 93 year time period, from 1834 to 1927. The publication dates of the identified requested documents were collated and are displayed in *Figure 6-19*.

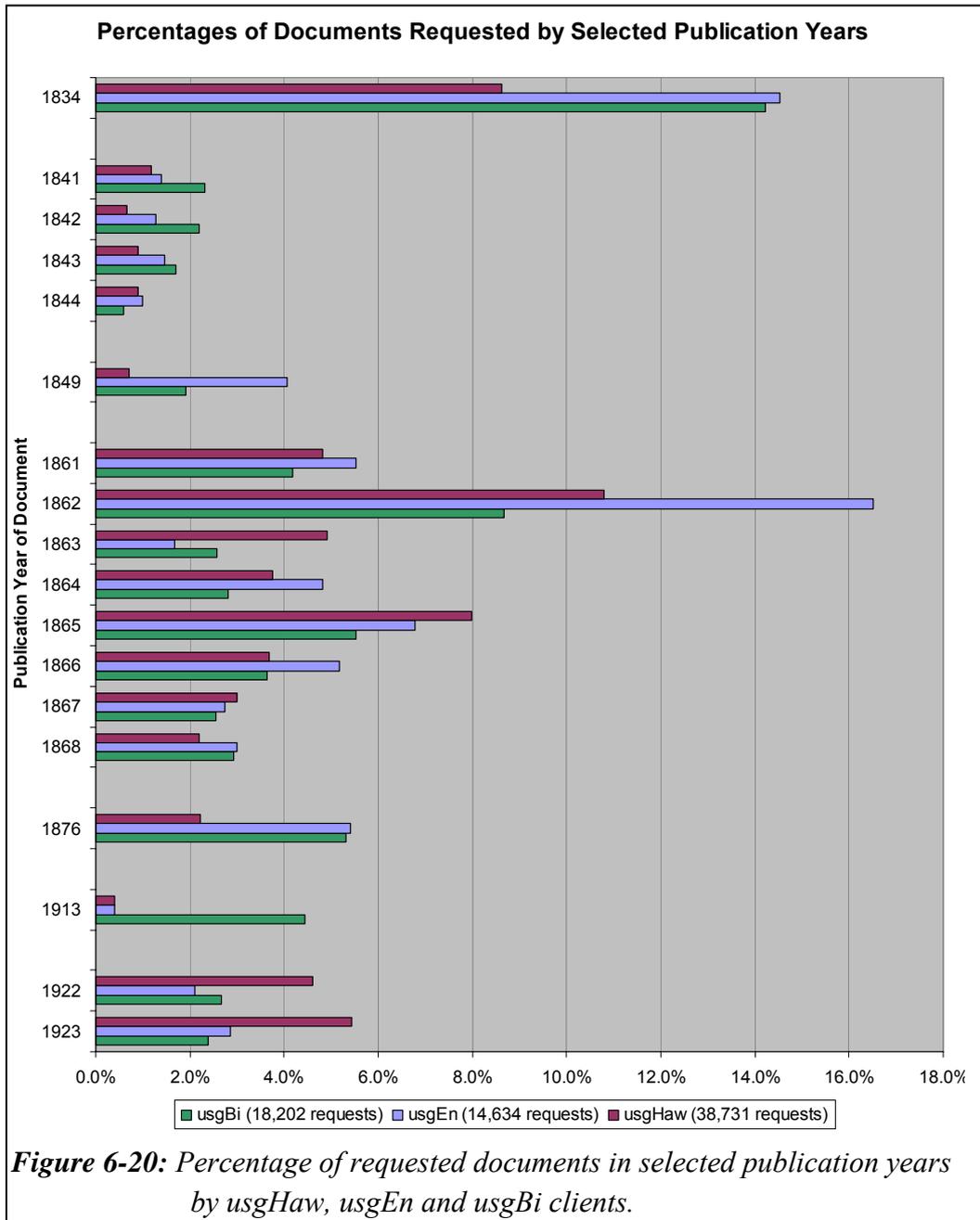
Figure 6-19 shows that the publication dates of the identified documents were most popular in the following years; 1834, 1841–1844, 1849, 1861–1868, 1876, 1913 and 1922–1923. These dates have been listed again in *Figure 6-20* and have been separated into years requested by usgHaw, usgEn and usgBi clients.

Figure 6-20 shows that 1862, 1834 and 1865 were the publication years most favoured by usgHaw clients. Nūpepa documents published in 1862 and 1834 were favoured even more by usgEn and usgBi clients. The documents published in 1865 were most favoured by the usgHaw clients.

Figure 6-20 does show that some publication years were favoured more by usgHaw clients than by other clients. For example 1922, 1923, 1863 and 1865. Other publication years, like 1849 and 1862, were more favoured by usgEn clients while others still, like 1842–1843 and 1913, were more favoured by usgBi clients. The reasons why certain client types displayed higher preferences for request documents in certain years could be due a number of factors; because of the content that was available in those publication years, the different methods clients



used to access documents or because of historic events that had different relevance to clients who used the Nūpepa digital library. Further research is needed to determine exactly why certain client types prefer Nūpepa documents written in certain years.



6.5.5 Documents With Associated Extracted Text Files

Nūpepa documents with associated extracted text files were likely to be requested more often than documents without extracted text files. Having an extracted text file allowed a document to be discovered when a query was submitted.

Of the pages that were able to be identified, 43,964 (61.4%) were Nūpepa documents that had an associated extracted text file (and could be subsequently searched on) and 27,603 (38.6%) were Nūpepa documents that did not have an associated text file. Counts of documents with associated extracted text files requested by usgHaw clients are displayed in *Table 6-9* along with counts of documents requested by usgEn and usgBi clients.

	usgHaw		usgEn		usgBi	
Total	38,731	100%	14,634	100%	18,202	100%
With Text	19,617	50.6%	12,040	82.3%	12,307	67.6%
Without Text	19,114	49.4%	2,594	17.7%	5,895	32.4%

Table 6-9: *The number of requests for documents with and without associated extracted text files by clients, usgEn and usgBi clients.*

Half of the documents requested by usgHaw clients had associated extracted text files. As the documents with associated extracted text files only represented 6.4% of the total Nūpepa documents, it can be stated that clients who used the Hawaiian language (usgHaw) favoured documents with associated extracted texts.

UsgEn clients had a much higher preference (82.3%) for requesting documents with associated extracted text files. This high preference shown by usgEn clients was a result of usgEn clients' dependency on the query facility to locate documents. This dependency meant that 82.3% of the documents requested by usgEn clients were sourced from only 6.4% of the total Nūpepa content.

6.5.6 Was the Detach Option Used?

Nūpepa clients had the option to detach a document. Selecting the detach option opened the document in a new browser window. Having two browser windows available meant that the client could retain a particular document view in one window while continuing to browse/query the Nūpepa digital library in the other window. The detach option was available for the extracted text, the preview image and the mid-size image views but was not available when viewing a full size image.

Data from the Nūpepa filtered log showed 1,989 documents were requested using the detach option in 2005. This number represents only 2.1% of the total documents requested. The detach feature was rarely used by usgHaw clients, only 693 documents were detached by usgHaw clients, a figure that represents only 1.4% of their total document requests.

The detach facility was used most by usgEn clients who requested 792 documents to be detached which still only represented 4.0% of the total usgEn documents requested. The usgBi clients requested 537 detachments at 2.3% of the total usgBi documents requested.

6.5.7 How Long to Read Documents?

Apparent document read times are used to indicate how long clients took to read actual documents (as opposed to interface pages) of the Nūpepa digital library. The average apparent document read times by usgHaw, usgEn and usgBi clients were generated from TLA and are displayed in *Table 6-10*.

(all times in seconds)	usgHaw	usgEn	usgBi	Total	Count
Full Size Image	169.0	133.3	134.2	147.6	4,329
Extracted Text	85.7	87.3	65.0	80.6	37,179
Mid-size Image	73.1	92.9	69.3	75.1	16,439
Preview Image	13.1	19.4	14.3	14.4	29,082

Table 6-10: Average apparent read times of document types by usgHaw, usgEn and usgBi clients.

Table 6-10 shows that, on average, usgHaw clients spent the longest time reading full size images, at almost 3 minutes. This was twice the average time taken to read extracted text documents (85.7 seconds) and mid-size images (73.1 seconds), and thirteen times the average time taken to view the preview images (13.1 seconds). UsgEn and usgBi clients also took the longest times reading full size images and the least times reading preview images.

6.5.8 Summary—Document Access Behaviours

This section analysed how Hawaiian language clients in usage sessions accessed documents from the Nūpepa digital library. Statistics drawn from TLA examined seven document retrieval behaviours exhibited by usgHaw clients.

The statistics revealed that selecting a document from a series list was method favoured by usgHaw clients. This preference contrasted with behaviours displayed by usgEn and usgBi clients of the Nūpepa digital library, and all clients of the Māori Niupepa, who preferred the query facility. The preference for usgHaw clients to use the series list to access documents indicates their awareness that only a small percentage (6.4%) of the documents had extracted text files, and also suggests they had prior knowledge of what documents were available in the Hawaiian Nūpepa digital library.

Statistics from TLA also revealed that preview images and extracted texts were the formats most preferred by usgHaw clients, although mid-size images were also significantly requested. The periodical *Ka Nupepa Kuokoa* was requested in three of four requests by usgHaw. The publication years 1862, 1834 and 1865 were most favoured by usgHaw clients. Half of the documents requested by usgHaw clients had associated extracted text files, an indication that the OCR process is being undertaken on documents relevant to Hawaiian language clients. The detach feature was rarely used by usgHaw clients. Apparent read times indicated that reading full size images took the longest time by usgHaw clients; approximately twice the time of reading mid-size images and extracted text files and thirteen times the time it took to view preview images.

This brief summary concludes the examination into how usgHaw clients accessed Nūpepa documents. The next stage, beginning in the next section, is to examine how queries were submitted by usgHaw clients.

6.6 Query Behaviours

This section analyses the query behaviour of usgHaw clients using the Nūpepa digital library. A query is the action taken by a client who has entered a term or terms into the input box of the Nūpepa search engine and then clicked the ‘Huli/Begin Search’ button.

The query statistics discussed in this section have been generated by the TLA process. Nine query behaviours are analysed in this section:

- the number of queries submitted in usgHaw sessions
- the number of query terms used in usgHaw queries
- the topics queried by usgHaw clients
- the language of query terms
- the altering of query parameters in usgHaw queries
- the use of query operators in usgHaw queries
- the number of documents requested following a usgHaw query
- the formats favoured following a usgHaw query
- the apparent document read times following a usgHaw query.

These nine query behaviours comprise the next nine sub-sections.

6.6.1 How Many Queries Were Submitted?

The number of queries submitted per usgHaw session was grouped into three categories; no queries per session, 1–2 queries per session and 3 or more queries per session. Percentages of queries in each category are displayed in *Table 6-11*.

Almost half of all usgHaw sessions did not involve a query, almost a quarter involved the use of one or two queries per session, while the remaining quarter involved the submission of three or more queries per session.

	usgHaw	usgEn	usgBi	all
Total Sessions	4182	3482	1090	8754
Sessions with no queries	49.0%	35.0%	29.6%	41.0%
1–2 queries (per session)	23.4%	29.4%	13.5%	24.6%
3+ queries (per session)	27.5%	35.7%	56.9%	34.4%

Table 6-11: *Queries per session in usgHaw, usgEn and usgBi sessions.*

The usgBi clients showed the highest preference for undertaking multiple queries per session, while the usgEn clients showed more use of queries per session than usgHaw clients. The numbers of queries per session submitted by the different client types were plotted and are displayed in *Figure 6-21*.

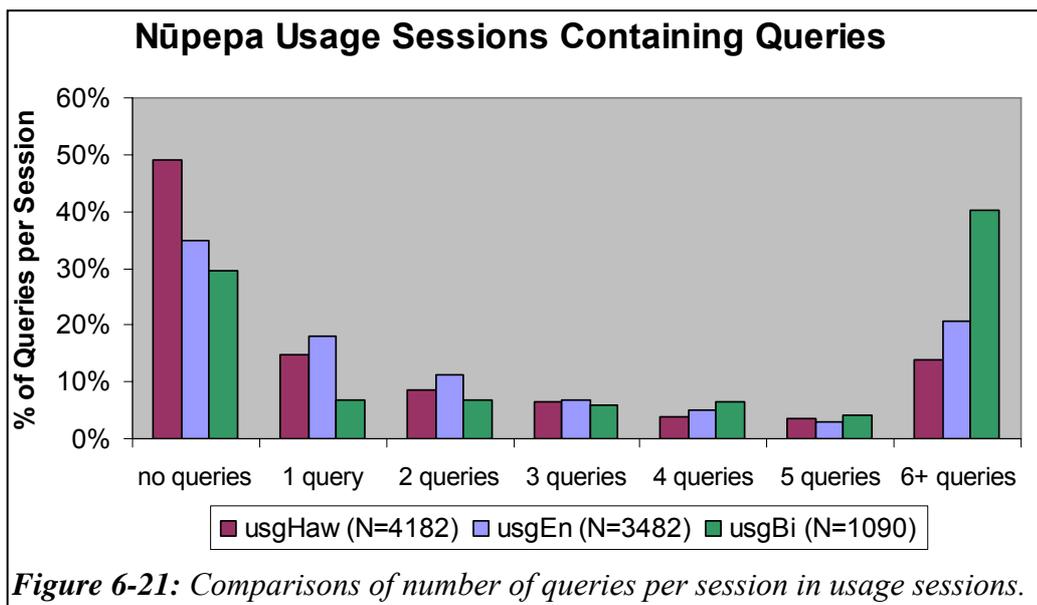


Figure 6-21 indicates again that usgHaw clients undertook the fewest number of queries per session. This figure also highlights that usgEn clients undertook the most single and double queries per session and usgBi clients undertook the most multiple queries per session.

6.6.2 How Many Query Terms Were Submitted?

The number of query terms submitted with each usgHaw query string was counted. Almost two thirds of all usgHaw queries involved just one query term. Approximately one quarter of all usgHaw queries involved two query terms. The remainder involved three or more query terms. The numbers and percentages of query terms per usgHaw query are displayed in *Table 6-12*, along with the numbers of query terms per usgEn and usgBi query.

	usgHaw		usgEn		usgBi	
Total	11,232	100.0%	14,966	100.0%	9,428	100.0%
1 term/query	7022	62.5%	10081	67.4%	5955	63.2%
2 terms/query	2580	23.0%	3553	23.7%	2484	26.3%
3 + terms/query	1630	14.5%	1332	8.9%	989	10.5%

Table 6-12: The number of query terms submitted per query by usgHaw, usgEn and usgBi clients.

In answer to the question posed, most of the time (62.5%), usgHaw queries involved the use of just one query term.

The percentages appear very similar for the three usage session types. However, chi square analysis returns a p value of $6.5 * 10^{-51}$, indicating that there is a significant statistical difference. The usgHaw clients displayed the highest preference for submitting queries with three or more terms (14.5%). The usgBi clients displayed the highest preference for submitting queries with two terms, and the usgEn clients displayed the highest preference for queries with just one term (67.4%).

6.6.3 What Was The Language Of The Query Terms?

The query terms submitted to the Hawaiian Nūpepa search engine were collated. There were 18,478 query terms used in 11,232 queries undertaken by usgHaw clients, 22,026 query terms used in 14,966 queries undertaken by usgEn clients and 14,342 query terms used in 9,428 queries undertaken by usgBi clients.

The most common terms queried are displayed in *Table B-1* which is located in *Appendix B*.

The language of the query terms was noted for the top 100 query terms submitted by usgHaw, usgEn and usgBi clients. Two languages were used, Hawaiian and English. The percentages of Hawaiian and English query terms have been plotted in *Figure 6-22*.

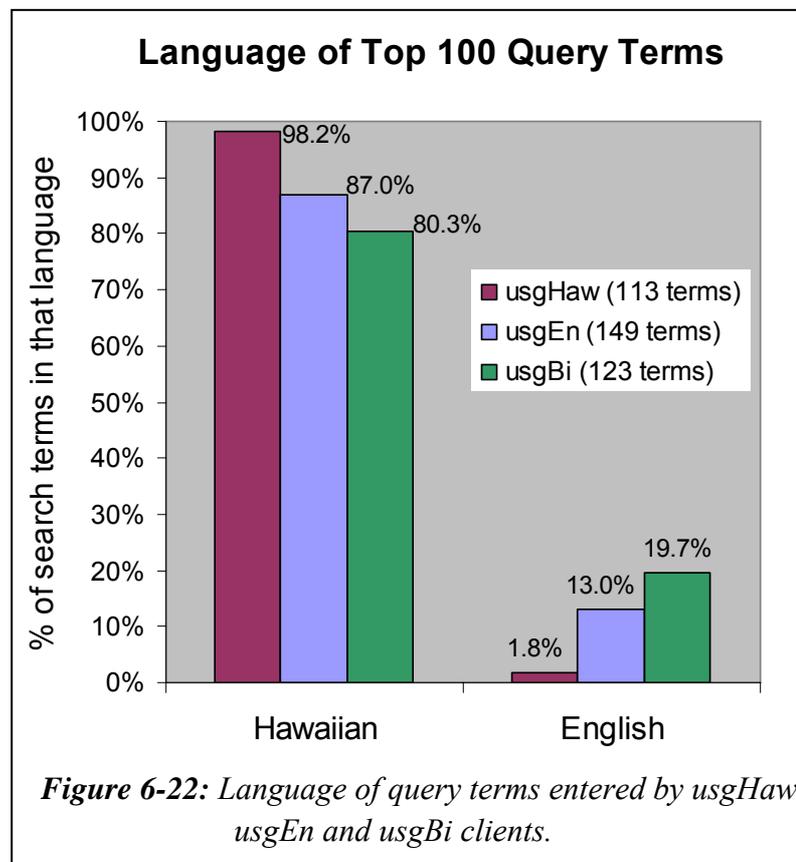


Figure 6-22 shows that Hawaiian was the predominant language of the query terms used in usgHaw queries. There was only one query term in the top 113 usgHaw query terms not in the Hawaiian language; a query for the term ‘*bush*’.

Considering that there are very few English language texts in the Hawaiian Nūpepa digital library it is surprising to see usgEn and usgBi clients using English query terms. The higher percentage of query terms in English by usgBi clients suggests that using English was a deliberate strategy undertaken by usgBi clients to assist in the information retrieval process.

6.6.4 Were Query Parameters Altered?

The Hawaiian Nūpepa digital library offers the user the ability to alter only one query parameter the Some/All parameter. This parameter gives the client the ability to query on *Some* of the query terms or *All* of the query terms. This parameter is available at the centre of the query page as shown in *Figure 6-2*.

The Some/All query parameter was not altered often by usgHaw clients; only 93 (3.2%) unique usgHaw users altered this parameter in 1.4% of the usgHaw queries.

A similar proportion of usgBi clients altered this parameter; 35 (3.3%) unique usgBi users in 0.8% of the usgBi queries. However, even fewer usgEn clients altered the Some/All parameter; 58 (1.9%) unique usgEn clients altered this parameter in 0.7% of the usgEn queries submitted.

6.6.5 Were Query Operators Used?

Counts from the web log indicate that Boolean operators (and, or, not, *, +, -) were not often used in query strings. Less than 0.9% of queries undertaken in usgHaw sessions involved the use of one of these operators, less than 0.3% in usgEn sessions, and less than 0.4% in usgBi sessions.

However, there was significant use of quotation marks to define precise phrase searching. Of the 11,232 usgHaw queries submitted 958 (8.5%) involved the use of quotation marks to specify exact phrase matching. This rate was higher than in usgEn queries (1.6%) and usgBi queries (2.5%). UsgHaw clients entered more query terms when undertaking queries (see §6.6.2); it is probable that those clients would be more likely to use an operator such as quotation marks to assist in defining results in multiple term queries.

6.6.6 Documents Requested Following a Query

Statistics derived from TLA on the Nūpepa log revealed how many documents were requested by usgHaw clients as a direct result of a query. There were 11,232 queries submitted in usgHaw sessions directly leading to the request of 10,702 documents; a return rate of 0.95 documents per usgHaw query. An average of 0.75 documents was requested by usgEn clients following a usgEn query and 1.16 documents requested by usgBi clients following a usgBi query.

The number of Hawaiian Nūpepa documents requested per usgHaw query was almost four times fewer than the number of Māori Niupepa documents requested per usgMi query. This can be attributed to the difference in the Nūpepa/Niupepa results pages. The Hawaiian Nūpepa results page, see *Figure 6-2*, displays two preview lines from the returned document to indicate the context in which the query term is used. In contrast, the Māori Nūpepa results page, *Figure 3-3*, does not show any preview lines but instead just lists the documents in which the query term is located. Clients of the Māori Niupepa collection must open (request) the documents to determine if the returned document is relevant to the query. Opening the documents to determine relevance results in a higher number of Māori Niupepa documents being requested per query as compared to the number of Hawaiian Nūpepa document requests per query.

If the number of Hawaiian Nūpepa documents requested following a query is fewer than the number of Hawaiian queries submitted, it could be an indication that the Nūpepa queries are not returning relevant results to the clients. Given this

scenario, the figures listed above suggest that queries submitted by usgEn clients are returning the least relevant results and queries submitted by usgBi clients are returning the most relevant results. The queries submitted by usgHaw clients, at 0.95 document requests per query, appear to be returning more relevant results than usgEn clients but less relevant results than usgBi clients.

6.6.7 Formats Favoured Following a Query

The document formats requested by usgHaw clients following a query were compared with the document formats requested when a query was not submitted. The percentages are displayed in *Table 6-13*, along with the figures for query and non-query document format requests by usgEn and usgBi clients.

	usgHaw		usgEn		usgBi	
	Non Query	After Query	Non Query	After Query	Non Query	After Query
Total	37,089	10,702	8,481	11,173	12,711	10,957
Extracted Text	21.7%	79.0%	46.0%	78.0%	28.3%	60.5%
Full Size Image	4.2%	4.5%	3.1%	2.5%	9.3%	12.1%
Mid-size Image	34.4%	3.5%	19.4%	3.9%	10.5%	4.9%
Preview Image	39.7%	12.9%	31.6%	15.6%	51.8%	22.5%

Table 6-13: Comparing types of documents requested not involving a query with those requested following a query from usgHaw, usgEn and usgBi clients.

Table 6-13 clearly shows that formats favoured when a query was not submitted were quite different from those favoured following the submission of a query. When a query had not been submitted almost 75% of the requests from usgHaw clients were for mid-size and preview images. However, when a query was submitted over 75% of the requests were for the extracted text documents.

UsgEn and usgBi clients displayed similar behaviours. When a query was not submitted the mid-size and preview images were preferred; when a query was submitted the extracted text documents were preferred.

6.6.8 Document Read Times Following a Query

The apparent read times of documents following a query were compared with the apparent read times of documents when no query was involved. The non-query and query average apparent read times by usgHaw clients are displayed in *Table 6-14*, along with the non-query and query average apparent read times by usgEn and usgBi clients.

(times in seconds)	usgHaw		usgEn		usgBi	
	Non Query	After Query	Non Query	After Query	Non Query	After Query
All Documents	49.7	96.9	56.3	85.6	35.0	73.6
Extracted Text	66.2	104.8	61.5	98.8	37.8	79.8
Full Size Image	167.2	174.8	148.9	120.3	123.4	144.0
Mid-size Image	71.9	115.4	95.8	81.9	60.2	91.7
Preview Image	40.5	23.1	51.6	18.8	25.5	17.1

Table 6-14: Comparing apparent read times of documents not involving a query with read times following a query for usgHaw, usgEn and usgBi clients.

Table 6-14 shows that, overall, the apparent read times of documents by usgHaw clients did change following the submission of a query. Apparent read times by usgHaw clients of extracted text documents, of full size images and of mid-size images were all longer following the submission of a query. However, the apparent read times of preview images by usgHaw clients were shorter following the submission of a query.

The apparent read times of documents by usgEn and usgBi clients also changed significantly following the submission of a query and were similar to those changes displayed by usgHaw clients.

6.6.9 Summary—Query Behaviours

This section analysed how queries were undertaken on the Nūpepa digital library by clients who used the Hawaiian language. Statistics drawn from TLA examined nine query behaviours exhibited by usgHaw clients.

The statistics revealed that almost half (49%) of all usgHaw sessions did not involve a query. When a query was submitted by usgHaw clients, half of them involved the use of one or two queries per session while the other half involved the submission of three or more queries per session. Where sessions contained a query, more than half (62.5%) of the usgHaw queries contained just one query term. While only 14.5% of usgHaw queries contained three or more query terms, this did represent a much higher preference by usgHaw clients for using multiple query terms than was shown by usgEn and usgBi clients. UsgHaw clients predominantly used Hawaiian language query terms (for a listing of the query terms submitted see *Appendix B*). The Some/All query parameter was altered only rarely by usgHaw clients, which was also the case in the use of Boolean operators in usgHaw queries. However, quotation marks were used in 8.5% of the usgHaw queries to define precise phrase matching.

Query statistics also revealed that usgHaw clients requested an average of 0.95 documents for every query submitted. UsgHaw clients who submitted queries primarily requested extracted text documents, while those who did not submit queries primarily requested mid-size and preview images. The apparent read times of documents by usgHaw clients increased for all document types following a query, apart from preview documents. The apparent read times of preview documents by usgHaw clients was much shorter following the submission of a query.

Usage of the Hawaiian language by clients who actively *used* the Nūpepa digital library has been discussed in three parts; usgHaw interface behaviours, usgHaw document access behaviours and the section summarised here, usgHaw query behaviours. Two further Nūpepa usage questions are posed: how did indigenous Hawaiian language usage differ from non-indigenous English language usage of the Nūpepa digital library; and how was the Hawaiian language bilingually used with the English language by clients of the Nūpepa digital library? These two questions will be addressed in the next two sections.

6.7 Hawaiian Language Usage/English Language Usage

The purpose of this chapter has been to define how the Hawaiian language was used by clients of the Nūpepa digital library in 2005. There is also merit in examining how this usage compares with English language usage of the Nūpepa digital library. If differences can be identified then they can be incorporated into digital library design to further assist indigenous language use, and non-indigenous language use, of a digital library.

A comprehensive usgHaw and usgEn comparison can be undertaken by examining the statistics listed in each sub-section of sections 6.4–6.6. The intention of this section, §6.7, is to summarise the key differences and similarities between how the Nūpepa digital library was used in Hawaiian and how it was used in English. This section comprises two parts; behaviour differences in expHaw and expEn exploratory sessions and behaviour differences in usgHaw and usgEn usage sessions.

Behaviour Differences in Exploratory Sessions

Nūpepa exploratory sessions in Hawaiian were slightly, but significantly, longer than exploratory sessions in English. The expHaw sessions had an average longer

duration (1.63 minutes compared with 1.56 minutes) and a higher average number of requests per session (3.0 requests compared with 2.3 requests) than expEn requests. The apparent interface page read times of expHaw clients was statistically similar to those by expEn clients. These figures imply that expHaw clients often requested one more interface page than the expEn clients before deciding not to use Nūpepa digital library.

Behaviour Differences in Usage Sessions

The Hawaiian Nūpepa digital library was used more by usgHaw clients than by usgEn clients. UsgHaw clients were responsible for 45.8% of the total usage requests while usgEn clients were responsible for only 28.8% of the requests (25.4% were submitted by usgBi clients). On average, usage sessions in Hawaiian were longer than those in English. The usgHaw sessions were longer in duration (16 minutes compared with 12 minutes) than the usgEn sessions and involved a larger average number of requests (22 requests compared with 16 requests).

UsgHaw clients took a similar time to read the interface pages (in usage sessions) as usgEn clients. This similarity was also noted in the exploratory sessions. The usgHaw clients had a higher percentage of return sessions (12.9% compared with 8.7%) and submitted more requests per return client (329 requests compared with 78 requests) than usgEn clients. A higher rate of return sessions and return session requests implies that the Nūpepa digital library was more relevant to usgHaw clients than to usgEn clients. Just over 93% of the usgHaw requests originated from within the Hawaiian Islands compared with only 62% of the usgEn requests.

It is difficult to compare usgHaw and usgEn language switching behaviours as the default language of the Nūpepa digital library was set to Hawaiian for all of 2005. Thus, many usgHaw clients did not need to switch the interface language while in theory, all of the usgEn clients did need to. Further, if usgHaw clients did entertain some form of language switching then they were often redefined as usgBi clients.

UsgHaw clients preferred to request documents from a series list, usgEn clients preferred to request documents by undertaking a query. This, in turn meant that usgHaw clients had a higher preference for requesting the mid-size or preview image formats whereas usgEn clients had a higher preference for requesting the extracted text formats. As usgEn clients were requesting a higher proportion of extracted text documents they were also requesting documents that mostly had associated extracted text files. Documents with associated text files were requested in 82.3% of the documents requested by usgEn clients compared with only 50.6% of the documents requested by usgHaw clients.

There appeared to be a small but significant difference in the periodicals and years of publication requested by usgHaw and usgEn clients. UsgHaw clients showed higher preferences than usgEn clients for requesting documents from *Ka Nupepa Kuakoa* and *Ka Hoku o ka Pakipika* and for the years 1863, 1865 and 1922–1923. Alternatively usgEn clients showed higher preferences than usgHaw clients to request documents from the *Ka Lama Hawai‘i* and documents in the years 1832, 1849, 1862 and 1876.

As mentioned above, usgEn clients preferred to use a query to access documents. This occurred even though only 6.4% of the Nūpepa documents were available to the search facility. UsgEn clients undertook more queries per session than usgHaw clients; however, these queries usually involved fewer query terms per query. UsgEn clients used English query terms in a small percentage of queries (13.0%) whereas usgHaw clients undertook almost no queries in English (1.8%). UsgEn clients, like usgHaw clients, made little use of query parameters. While usgHaw clients used quotation marks to specify exact phrase matching in 8.5% of queries, usgEn clients only used this operator in 1.6% of queries. As usgHaw clients showed a higher propensity to submit queries with multiple terms it also follows that usgHaw clients would make more use of this phrase matching operator.

UsgEn queries did not lead to as many document requests as usgHaw queries; there was an average of only 0.75 documents requested following a usgEn query compared with 0.95 documents requested following a usgHaw query. When the

rate of document requests per query is lower than one it suggests that many queries are not returning relevant results.

The submission of a query affected the types of documents that were requested. When a query was not submitted usgHaw clients preferred to request full size images. However, following a query usgHaw clients had a much higher preference for extracted text documents, while the preferences for full size and preview images reduced by almost a half. UsgEn clients did not show the same differences following a query. UsgEn clients had the highest preference for extracted text documents whether a query was submitted or not. The only difference following a query by usgEn clients was a reduction in the preference for accessing preview images which subsequently led to an increase in the preference for extracted text documents.

The submission of a query also affected the apparent read times of documents. In usgHaw sessions the apparent read times following a query decreased for full size images, remained the same for extracted text documents but increased for preview images. In usgEn sessions, the apparent read times increased following a query increased for full size images, decreased for extracted text documents and stayed the same for preview images.

UsgEn clients showed a strong preference for switching the interface language from Hawaiian to English usually at the start of a session. When usgHaw clients switched the interface language from Hawaiian to English they did it towards the end of a session. Both client types were recorded switching the interface language from English to Hawaiian, though not to the same extent, and both client types generally made this switch at the beginning or end of a session.

6.8 Bilingual Usage

The analysis of the Māori Niupepa digital library has suggested that bilingual use of the Niupepa digital library was a strategy undertaken by a small number of clients to increase the effectiveness of information retrieval (see §3.8). Statistics generated from TLA of the Hawaiian Nūpepa digital library indicate that this strategy was undertaken by a large rather than small number of clients. There were 1,090 unique Nūpepa clients categorised as usgBi clients and these clients were responsible for 25.4% of the requests to the Nūpepa collection. There were almost as many requests undertaken by usgBi clients (50,359 requests) as by usgEn clients (57,154 requests).

By definition (see §2.4.3) usgBi sessions must contain at least 10 requests. Perhaps because of this, usgBi sessions had almost twice the duration and twice the number of requests as usgHaw and usgEn sessions. Likewise, the number of documents requested per session was much higher for the usgBi sessions; 22.2 documents per session while usgHaw sessions requested 11.6 documents per session and usgEn sessions requested 5.9 documents per session.

There were similar proportions of return usgBi and usgEn clients at 13% of the total clients. However the return usgBi clients were only responsible for 17% of the total usgBi requests whereas the return usgHaw clients were responsible for 64% of the total usgHaw requests.

Document access behaviours exhibited by usgBi clients tended to fall between those behaviours displayed by usgHaw clients and those behaviours displayed by usgEn clients. The methods used to access documents, the preferred document formats, the favoured periodicals and publication years, the preference for documents with extracted texts, the use of the detach facility and the apparent read times by usgBi clients were all behaviours that fell between usgHaw and usgEn client behaviour.

The manner in which usgBi queries were undertaken was different from other clients. UsgBi clients undertook more queries per session (8.7) than usgHaw (2.7) or usgEn clients (4.3). UsgBi clients also used more English in query terms (19.7%) than usgHaw (1.8%) or even usgEn clients (13.0%). UsgBi clients appeared to have the most effective querying strategies as they recorded the highest rate of document requests (1.2) following a query. UsgBi clients were also the least interested in accessing the extracted text formats following a query. While usgBi clients did request extracted text formats in 60% of the documents requested following a query, this percentage was 20% lower than requests by usgHaw and usgEn clients. UsgBi clients seemed more interested than usgHaw or usgEn clients in the image form of Nūpepa documents following a query.

In summary, it appears that the Hawaiian language was used with the English language in a deliberate strategy to assist in information retrieval and to request a relatively high number of documents. The way the indigenous language was used bilingually in the Hawaiian Nūpepa digital library is remarkably similar to the way the indigenous language was used bilingually in the Māori Niupepa digital library. The striking difference was that the Hawaiian Nūpepa digital library was used considerably more bilingually (25.4% of requests) than the Māori Niupepa digital library (2.3%).

6.9 Summary

Statistics from Transaction Log Analysis (TLA) have been used in this chapter, to determine how the Hawaiian language was used by clients of the Nūpepa digital library in 2005. This chapter began by giving background information about the Hawaiian Nūpepa digital library and then described, with the aid of screenshots, how clients could access documents within the digital library.

Requests to the Nūpepa digital library were then categorised into sessions and language of sessions to enable investigation whether the Hawaiian language was used by clients of the Nūpepa digital library. Results indicated that the Hawaiian Nūpepa digital library was used predominantly in the Hawaiian language. Almost half of the usage requests were defined as Hawaiian language usage sessions (usgHaw), a little over one quarter were defined as English language usage sessions (usgEn) and the remaining quarter were defined as sessions where the clients used the interface in both Hawaiian and English (usgBi).

Analysis was undertaken to determine how Hawaiian language clients explored the Nūpepa interface but then did not go on to use it. Exploratory requests comprised only 1.2% of the total requests to the Nūpepa digital library. Hawaiian language exploratory clients, in a similar manner to English language exploratory clients, only made two or three requests per session and most sessions were completed in less than half a minute.

Analysis of how Hawaiian language clients actually used the Nūpepa digital library was undertaken using statistics generated from TLA. These statistics showed that usage sessions undertaken by usgHaw clients averaged 16 minutes and 22 requests. The Hawaiian language usage analysis was discussed in three categories; usgHaw interface behaviours, usgHaw document access behaviours and usgHaw query behaviours.

UsgHaw interface behaviours showed the average apparent read time of an interface page by usgHaw clients was 38 seconds. At least 13% of the usgHaw

sessions were return sessions and usgHaw clients who returned to use the Nūpepa digital library were responsible for almost 64% of the total usgHaw requests. Fourteen out of fifteen requests submitted by usgHaw clients originated from the Hawaiian Islands. UsgHaw clients did not switch the language of the interface often; less than 1 usgHaw session in 12 involved an interface language switch.

UsgHaw document access behaviours indicated that usgHaw clients preferred to select documents from a series list. This preference is contrary to behaviours displayed by usgEn and usgBi clients of the Nūpepa digital library, and by all clients of the Māori Niupepa digital library, who all showed the highest preference for selecting documents by using the query facility. Statistics also revealed that preview images and extracted texts were the formats most preferred by usgHaw clients, although mid-size images were also significantly requested. Half of the documents requested by usgHaw clients had associated extracted text files. The detach feature was not used significantly by usgHaw clients. The average apparent document read time by usgHaw clients were almost three minutes for a full size image, almost 1.5 minutes for an extracted text file, just over 1.2 minutes for a mid-size image and 40 seconds for a preview image.

UsgHaw query behaviours revealed that almost half (49%) of all usgHaw sessions did not involve a query. When a query was submitted, half involved the use of 1–2 queries per session while the other half involved the submission of three or more queries per session. UsgHaw clients entered more query terms per query than usgEn and usgBi clients. UsgHaw clients predominantly queried terms written in the Hawaiian language. The Some/All query parameter was not used significantly by usgHaw clients; however quotation marks were used in 8.5% of the usgHaw queries to define precise phrase matching.

Query behaviours also revealed that usgHaw clients requested an average of 0.95 documents for every query submitted. UsgHaw clients who submitted queries primarily requested extracted text documents, while usgHaw clients who did not submit queries primarily requested mid-size and preview images. UsgHaw clients took longer to read all types of documents following a query, apart from the preview images.

Activities undertaken by Nūpepa clients who used Hawaiian were compared with activities undertaken by Nūpepa clients who preferred English. ExpHaw clients, on average, requested one more interface page per session than the expEn clients. UsgHaw sessions were a third longer than usgEn sessions in duration and in number of requests. More usgHaw clients than usgEn clients returned to use the Nūpepa digital library and the usgHaw clients were responsible for more requests. Almost all usgHaw requests originated from the Hawaiian Islands while a significant proportion (38%) of usgEn requests originated from outside of the Hawaiian region. UsgHaw clients preferred to select documents from the series list and favoured accessing preview and mid-size images. UsgEn clients preferred to select documents from the results of a query and favoured accessing extracted text documents. These preferences led to usgHaw and usgEn differences in document access behaviours and query behaviours (as has been discussed in §6.8).

The bilingual use in Hawaiian and English was analysed in the previous section. This section revealed that the usgBi clients were a significant Nūpepa client group responsible for a quarter of all requests to the Nūpepa digital library. UsgBi sessions were longer than usgHaw and usgEn sessions, both in duration and in number of requests. UsgBi clients demonstrated document access behaviours that were between those displayed by usgHaw and usgEn clients. However, the usgBi clients were the client group most focused on using the query facility to access documents. UsgBi clients exhibited unique searching behaviours, such as using a high percentage of English search terms, to achieve accurate results from queries.

This chapter has charted Hawaiian language usage by clients of the Nūpepa digital library in 2005 and has clearly shown how the Hawaiian language was used by clients to retrieve information from the Nūpepa digital library. Like te reo Māori, the Hawaiian language was being used by choice without prompting or coercion. This chapter has provided a second example of how an indigenous language was used by clients of a digital library.

The previous four chapters have examined *how* indigenous languages were used by clients of digital libraries. TLA has been used to give comprehensive analysis of what indigenous language usage occurred on two separate bilingual indigenous

digital libraries in 2005. The statistics have been used to give a picture of what happened but not why it happened. Chapter 7 will use two methods to gain insights into *why* indigenous languages were used by clients of digital libraries.

7 Why was the Indigenous Language Used by Digital Library Clients?

Statistics derived from Transaction Log Analysis were used in the previous four chapters to investigate *how* indigenous languages were used by clients of digital libraries. The results showed that use of indigenous languages did occur suggesting that digital libraries provide an electronic medium for the use of an indigenous language. Why did clients choose to use the indigenous language when working with the digital? The purpose of this chapter is to seek feedback from clients of digital libraries to suggest answers to this question.

Two methods were used to obtain client feedback; a web questionnaire and client interviews. The web questionnaire was placed on the Māori Niupepa web site for 250 days in 2006. The questionnaire was presented to clients when they requested the home page of the Niupepa digital library. There were 60 responses to the web questionnaire. Client interviews were undertaken with users or potential users of the Niupepa and Nūpepa digital libraries. Those interviewed comprised 3 clients who used the Māori Niupepa interface in te reo Māori, 8 clients who used the Hawaiian Nūpepa interface in Hawaiian and 19 clients who used the Māori Niupepa interface in English. Although the sample size was small some relevant responses were collected which give important insights into indigenous language usage of digital libraries.

The question of why indigenous languages were used by clients of digital libraries will be answered in two parts; why indigenous language speaking clients used

these particular digital libraries and why they used them in an indigenous language. Feedback indicates the content of the digital library attracted indigenous language clients, along with the fact that this content was in the indigenous language. The indigenous language was used by clients of digital libraries because the indigenous language matched the content of the digital library, was the default language of the digital library and because the clients wanted to use the indigenous language as often as possible.

7.1 Web Questionnaire Methodology

As mentioned in the introduction, two methods were used to obtain client feedback. This section describes the first method, the web questionnaire. Web questionnaires, also known as on-line questionnaires, are a widely used technique to solicit feedback from clients who by their very nature cannot be questioned directly. They are considered easier, more efficient and less error prone than paper based questionnaires (Bradburn, Sudman, & Wansink, 2004). A full description is given of the web questionnaire, followed by an explanation of how it was presented to Niupepa clients in 2006.

7.1.1 Web Questionnaire Description

The questionnaire specifically sought to determine the contexts in which the indigenous language was used. As it was only used on the Māori Niupepa web site, it collected information only about the contexts in which te reo Māori was used by Niupepa clients.

The web questionnaire is shown in *Figure 7-1*.

Web Questionnaire Form

Three Minute Niupepa Feedback

Tēnā koe. Tēnei te inoi atu ki a koe kia tautokohia mātou i a mātou e rangahau kōrero ana e pā ana ki te kohikohinga nei.
Please take a few minutes to answer some anonymous questions about your usage of this website.
For questions about this feedback form please email Te Taka Keegan (tetaka at cs.waikato.ac.nz). Ka nui te mihi.

Ka whakamahia e koe tēnei kohinga mō te aha? (pāwhiritia ngā whakautu e hāngai ana)
What do you use this collection for? (select which answers apply)

Rangahau kōrero e pā ana ki te iwi *Researching tribal information*
 Rangahau kōrero e pā ana ki a koe, ki tō whānau rānei *Researching personal or family information*
 Rangahau ngā āhuatanga o neherā *Researching historical practices*
 Rangahau kupu/āhuatanga o te reo *Research on words/language*
 Rapu kōrero whakamāori/whakapākehā *Search for translation texts*
 Whakangahau/rapu kōrero hei pānui noa *Fun/recreational reading*
 Ētehi atu, whakamōhio mai *Other, please specify*

He aha ngā tumomo mahi i tirohia ai e koe tenei kohikohinga?
For which types of industry are you accessing this collection?

Mahi mō te iwi *Tribal research*
 Rangahau i te whare wānanga *Academic research*
 Mahi ako *Teaching*
 Kawanatanga *Government*
 Ture *Law*
 Paoho *Media (tv, radio, print)*
 Rongoa *Medicine*
 Whaiaro motuhake *Personal research*
 Ētehi atu, whakamōhio mai *Other, please specify*

He aha ngā painga o te whakamahi i te kohinga nei?
What do you find most rewarding about using the collection?

He aha ngā mea kāore i te pai o te whakamahi i te kohinga nei?
What do you find most frustrating about using the collection?

I a koe e whakamahia ana te kohikohinga nei i te nuinga o te wā ka noho te reo o te atanga ki tēwhea reo?
As you are using this collection, the language of the interface is mostly set to which language?

Māori
 English
 Aua *Not sure*

He kōrero i tua atu?
Any other comments?

Done

Figure 7-1: A screen shot of the web questionnaire form

The web questionnaire was a parallel language form with texts written in te reo Māori first, followed by English equivalents in italics. The form began with a greeting in both te reo Māori and English, followed by six questions and a button to submit the form. The six questions asked were:

- *What do you use this collection for?*
- *For which types of industry are you accessing this collection?*
- *What do you find most rewarding about using the collection?*
- *What do you find most frustrating about using the collection?*
- *As you are using the collection, the language of the interface is mostly set to which language?*
- *Any other comments?*

One question not asked was “Why do you use the digital library in te reo Māori?” This was seen as a leading question that might bias the clients’ responses to being overtly positive or overtly negative towards te reo Māori.

The web questionnaire was titled the ‘Three Minute Niupepa Feedback’. It was given this title in the hope that clients would realise that answering the form would take only three minutes of their time and, thus, encourage more clients to complete it.

7.1.2 Web Questionnaire Presentation

The web questionnaire was designed to be as unobtrusive as possible to clients of the Niupepa web site. The questionnaire would only appear to those clients who had enabled cookies on their web browsers and requested the Niupepa home page for a second time. It was a pop-up window that could simply be ignored, and full access to the Niupepa digital library could continue. Once a client had completed and submitted the questionnaire it would not re-appear to that client.

Two cookies were created to handle the client interaction with the questionnaire. A cookie called *niupepasurvey_visited* and *niupepasurvey_blocked* were loaded onto a client’s computer when they first visited the Niupepa web site. Both

cookies were set to 'true'. The *niupepasurvey_visited* cookie was set to expire on the 1 January 2007, while the *niupepasurvey_blocked* cookie was set to expire when the client's current session with the Niupepa web site expired. When the client started a second session on the Niupepa web site the states of these cookies were examined; if the *niupepasurvey_visited* cookie was set to 'true' and the *niupepasurvey_blocked* cookie was set to 'false' then the questionnaire was displayed to the client. Once the client had submitted the survey the *niupepasurvey_blocked* cookie was set to 'true' with an expiry date of 1 January 2007. As both cookie states were now set to 'true' the questionnaire was not redisplayed to that client on subsequent visits to the Niupepa site.

The web questionnaire was activated on the Puka and Thor web servers on 20 of February 2006 and deactivated 3 November 2006. In total, the web questionnaire was active for exactly 250 days and sixty responses were received. Their responses are analysed in detail in §7.2.

7.2 Web Questionnaire Responses

Sixty responses to the web questionnaire were received. Some further responses were recorded when the questionnaire was being set up and tested; however, as these were not submitted by actual clients of the Niupepa digital library they were not included in the questionnaire analysis. There were 48 questionnaires logged by Puka (the national web server) and 12 by Thor (the international web server).

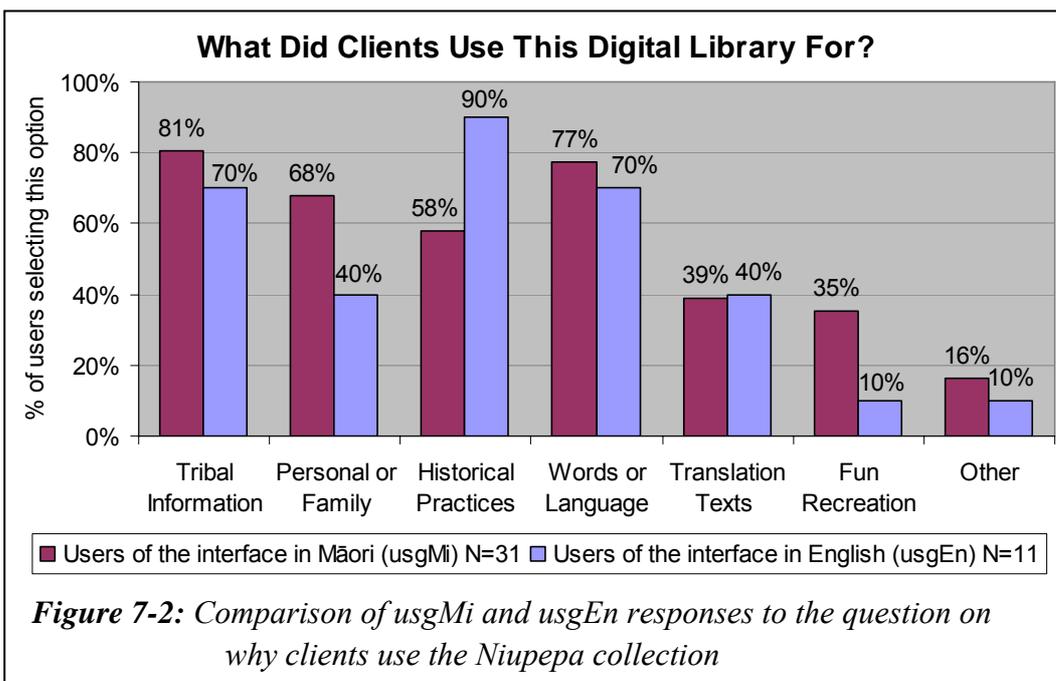
While sixty responses were recorded, not all of the responses included answers for all of the questions. There were ten nil responses, that is, responses that just had the time stamp recorded. A nil response is a result that would occur if the client scrolled through to the end of the form without answering any of the questions and clicked on the *Tukua/Submit* button. Two of the nil responses were from national responses; the other eight were from international responses. The international clients appeared to show a strong resistance to this method of collecting data whereas the national clients appeared prepared to respond.

Of the 60 clients that responded to the questionnaire, 31 stated that they used the interface in te reo Māori (usgMi), 11 that they used the interface in English (usgEn), 6 that they were unsure which language they used the interface in and the remaining 10 gave no indication of which interface language they preferred to use.

As shown in *Figure 7-1* the web questionnaire asked six questions. The responses received for each of these questions are detailed in the next six sub-sections. This section then concludes with a summary of the information deduced from the responses to the web questionnaire.

7.2.1 What Did Clients Use the Digital Library For?

The first question asked on the web questionnaire was, ‘*what do you use this collection for?*’ As shown in *Figure 7-1* clients could choose from six predetermined options or use the seventh option to specify an unlisted option. The percentages of responses selected for each client type have been graphed in *Figure 7-2*.



The Niupepa digital library was used by usgMi clients predominantly for *researching tribal information*, the option selected by 25 of 31 usgMi respondents. This response was closely followed by the *words or language* option which was selected by 24 of 31 usgMi respondents. *Researching personal or family* information also rated highly with 21 of 31 usgMi respondents.

It appears that usgMi clients had a stronger preference for using the Niupepa digital library in a wider range of categories than usgEn clients. Most apparent were the stronger preferences for the categories *fun/recreational reading* (35% compared with 10%) and *researching personal or family information* (68% compared with 40%). The usgEn clients displayed a higher preference for researching *historical practises* (90% compared with 58%).

Seven respondents selected the *other* option four of which entered what the other reasons were. These were entered as:

"Work Purposes"

"He mea tino ataahua, he mea koi ki te hinengaro"

[something very beautiful, something to sharpen the mind]ⁱ

"whakamohio au ko wai au i roto i tenei ao hurihuri"

[to know who I am in this changing world]

"Hei hanga rauemi"

[to create resources]

Figure 7-2 suggests that English language clients preferred to use the digital library to find information on historical practices, while te reo Māori clients preferred to locate tribal information, personal information and information about (reo Māori) words and language. However, there was a low response rate number; there were only 31 responses classified as usgMi and only 11 responses classified as usgEn. This count is too low for any firm conclusions to be drawn.

ⁱ All translations in this chapter are by the author.

7.2.2 Which Research Sectors Used the Digital Library?

The second question was about the research sectors (or industry types) that accessed this collection. As displayed in *Figure 7-1*, eight options were provided plus the option to select ‘*other*.’ The responses collected have been calculated as percentages and are displayed in *Figure 7-3*.

The usgMi respondents on average selected two responses to this question, while usgEn respondents generally selected three. This lower amount of responses selected suggests that the usgMi respondents were using the digital library, and possibly reo Māori resources in general, across fewer sectors than usgEn respondents.

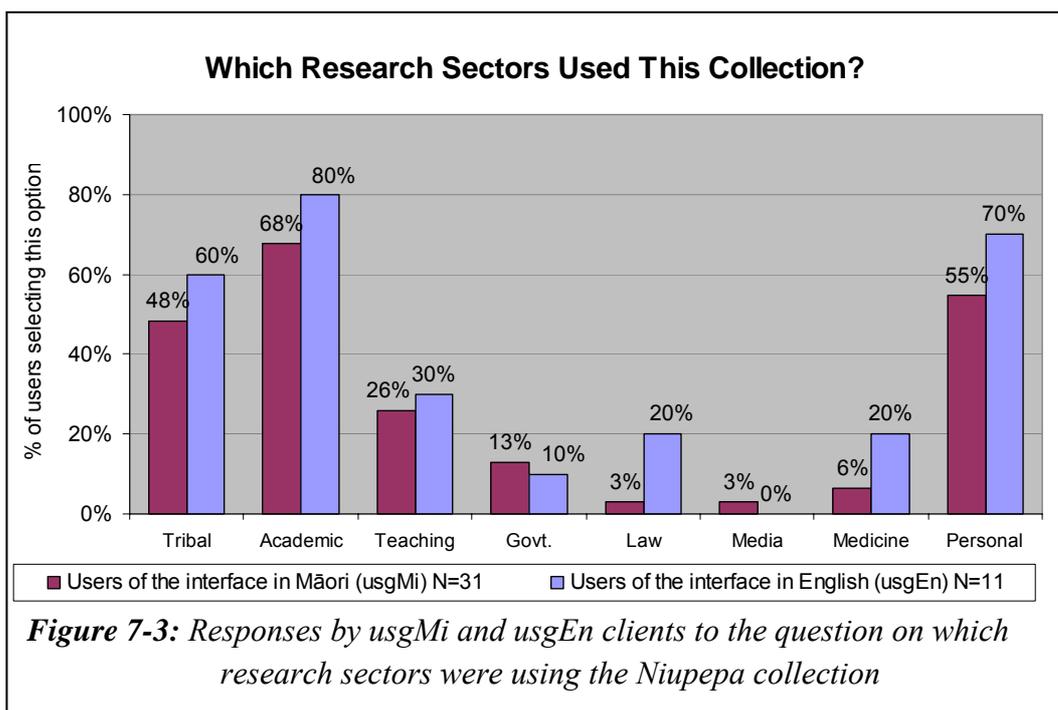


Figure 7-3 indicates that three research sectors were prominent users of the Niupepa digital library in te reo Māori. These fields were *academic* (21 of 31 responses), *personal* (17 of 31 responses) and *tribal* (15 of 31 responses). While it appears that the Niupepa digital library was not used often in te reo Māori by clients in other sectors, it was difficult to draw solid conclusions because of the low number of responses received to the questionnaire.

Four respondents selected the *other* option. These were entered as:

“Hei rauemi ki te whakapai i te Wikipedia”

[As a resource to enhance the Wikipedia]

“Mahi mo te whanau me te hapu”

[work for the family and sub-tribe]

“Treaty claims research”

“Assignment”

7.2.3 What Did Clients Find Most Rewarding?

There were 44 responses for the first open ended question, ‘*what did you find most rewarding about using the collection?*’ Of these, 31 respondents classified themselves as usgMi clients, 11 respondents classified themselves as usgEn clients and a further two respondents gave no indication.

Both usgMi clients (24 of 31) and usgEn clients (7 of 11) mentioned that the newspaper content of the Niupepa digital library was the most rewarding aspect of the digital library. A comment on the content was by far the most common response to this question. Some of the descriptions describing the value of the content were poetically written:

“Ko te mea nui, kua hikoi au i te wa e ora ai a tatou
kuia, a tatou koroua”

[The important thing is that I have walked in the times that my ancestors
were alive]

“Ko te hohonutanga o nga korero kai runga i nga parirau
o nga manu nei.”

[The deepness of the information delivered on the wings of these birds (the
newspapers were often referred to as messenger birds)]

“Te wairua o a tatou kuia, a tatou koroua i te wa e ora ai ratou, e tipu ake ratou e tangata whenua ana ratou nga mea o te ao maori hei taonga tuku iho mo tatou. E kore au e ngaro, he kakano i ruia mai i Rangaiatea.”

[The spirit of our ancestors in the times they were alive, when they grew up, when they were masters of the Māori world passed on as a treasured heirloom for us. I can never be lost, I am a seed sown from Rangiātea.]

UsgMi responses to the web questionnaire indicated that the second most rewarding aspect of the Niupepa digital library was te reo Māori. Te reo Māori was selected in 11 of 31 usgMi responses. Comments were made regarding two areas; te reo Māori written in the newspapers and te reo Māori used in the digital library interface. Examples of comments made in the web questionnaire include:

“Ko te maamaa me te ataahua o te reo o neheraa”

[The clarity and beauty of the old language]

“Firstly, reading the thoughts of my tipuna who wrote letters to the newspapers, through which I have gained an understanding of the choices that they made. Secondly improving not only my understanding of the Maori language but more importantly learning my own dialect as written by my own people!!”

The ability to undertake queries was the third aspect that clients found rewarding, as mentioned by six usgMi respondents and four usgEn respondents. Also mentioned was the availability of the information, having additional information (English abstracts and newspaper commentaries) and the browsing features offered on the web site. Two usgMi respondents also used this question to thank the builders of the digital library for making the Niupepa collection available.

7.2.4 What Did Clients Find Most Frustrating?

There were 42 responses recorded to the second open-ended question which asked, ‘*what did you find most frustrating about using the collection?*’ There were 28 respondents who classified themselves as usgMi clients, six as usgEn clients and a further eight respondents who gave no indication.

The most common response to this question was nothing, that is, there was nothing frustrating. This response was given by 10 (of 28) usgMi clients and 3 (of 6) usgEn respondents. Typical responses were:

“Horekau he raruraru”

[No problems]

“Ka pai te katoa, ki ahau.”

[It’s all good to me]

“Kare kau, he taonga, he waka hua [huia], he mea miharo, he mkea[mea] ataahua...”

[Nothing, a treasure, a repository of treasures, something amazing, something beautiful]

“...kotahi anake te raru, kaore e whai atu te haora, me te ra, me te wiki, me te marama, me te tau ki te noho ana ki te panui pukapuka (only one problem, not enough hours in the day, the week, the months or the year to sit down and feel the wairua o enei korero)”

[...to sit down and feel the spirit of these texts]

There were 18 responses from usgMi clients that did list difficulties. The most common, mentioned by 7 respondents was the fact that the OCR texts contained errors. There were also issues with navigation (2), searching (2), typing errors on the home page (2), the large number of results following a search (2), printing (1), the layout (1), the highlighting (1) and missing newspapers (1).

7.2.5 What Language Was Used by Clients Who Responded To the Questionnaire?

Analysis of this question is in two parts. First, which interface language the respondents stated they used, and second, which language the respondents actually used to answer the open-ended questions in the web questionnaire.

There were 48 responses to the question about which interface language clients used. There were 31 respondents (64.6%) who stated that they used the Niupepa digital library with the interface language set to reo Māori (usgMi), 11 (22.9%) who stated that the interface language was mostly set to English (usgEn) and 6 (12.5%) who stated that they were unsure which language the interface was set to.

For the open-ended questions 32 mostly used te reo Māori when answering the questionnaire. Of these, 27 stated that they used the reo Māori interface, one used the English interface and four seemed unaware of which interface they used as they chose the '*Aua, Not sure*' option. The client who responded in te reo Māori but stated they used the English interface was the only example of a potential reo Māori client choosing to use the interface in English. The respondent actually stated that this was the first time that he/she had used the collection. This is evidence to suggest that perhaps bilingual clients will use an interface in English until they become familiar with it and then use the interface in the indigenous language.

Three clients stated that they used the Niupepa interface in te reo Māori but responded to the web questionnaire in English which indicate that these clients were not highly literate in te reo Māori. This, in turn, could suggest that the Māori interface was still suitable for use by clients with little reo Māori knowledge.

It should also be remembered that while the web questionnaire was written in two languages, te reo Māori was the prominent language seen. It appeared on the first line and began each question. Te reo Māori was not written in italics. Perhaps if the more commonly spoken English language had been the prominent language

more responses may have been written in English, perhaps fewer nil responses and, possibly, more responses received over all.

7.2.6 Further Comments From the Web Questionnaire

The final question solicited further comments. This question was answered by 32 respondents; 23 responded in te reo Māori and 9 in English. The overwhelming response was positive; thanks and praise for the Niupepa digital library and encouragement for those who had made it available on the Web. There were 27 out of the 32 respondents who gave thanks, praise or encouragement. Some examples of these responses are listed below;

“E kore e mutu aku mihi ki a kotou {koutou} ki te rauemi nei me o taua nei tipuna nana nei te kupu i tuku ki te komata o te rangi hei tahu mo taku whare korero.”

[My unending thanks to you people and to this resource and to our ancestors who sent their words to the exalted heavens as a backbone for my house of learning]

“He tino taonga ngaa niupepa nei. Moo ngaa mea peenei i a au nei, horekau atu he huarahi e whiwhi ai au ki eenei koorero. Ka mihi kau atu.”

[These newspapers are a valuable treasure. For people like me, we have no other avenue to access this information. I extend my thanks.]

“Very valuable. Keep up the good work!”

Responses to the final question also included some suggestions and some advice. The suggestions and advice were either written separately or included with some responses expressing gratitude. These responses included a suggestion to use a PDF format for presenting the Niupepa documents, sourcing other newspapers to add to the collection, sourcing other material to add to the collection, correcting the syntax errors and a call not to translate the niupepa documents into English.

One final aspect noted was that usgMi clients appeared less likely than usgEn clients to state a problem with the digital library and more likely to praise. When a problem was stated by usgMi clients, often the English language was used.

7.2.7 Summary—Web Questionnaire Responses

Using the web questionnaire to elicit user feedback was an important exercise because it generated a wide range of responses from clients who used the Niupepa digital library in te reo Māori. The feedback raised some important issues about why clients have used the Niupepa digital library in te reo Māori, issues that otherwise may not have come to light.

It appears that te reo Māori was used for a wide range of reasons but across a relatively small range of industries. The primary reasons for accessing the digital library in te reo Māori were for undertaking tribal research, research on te reo Māori and/or research on personal or family information. A significant number of usgMi respondents also selected *researching historical practices*, *search for translation texts* and *fun/recreational reading* as reasons for accessing the digital library. The range of industries that appeared to be accessing the Niupepa digital library was limited to academic institutions, tribal organisations, with personal researchers also using the resource.

In contrast, the Niupepa digital library appeared to be used in English for fewer reasons by a wider range of industries. UsgEn clients were most likely to use the digital library to research historical practices and while interest was shown in other topics it was generally lower than that shown by usgMi clients. A wide range of industries appeared to be accessing the Niupepa digital library in English; academic institutions, tribal organisations, along with personal researchers were selected by more than half of the usgEn respondents but a significant number also selected the education (teaching) and law and health (medicine) industries.

Both usgMi and usgEn clients stated that the content in the historic newspapers was the most rewarding part of the Niupepa digital library. UsgMi users showed a

strong interest in the written reo Māori used in the newspapers themselves. UsgMi clients also indicated that they were pleased that the interface of the Niupepa digital library was in te reo Māori. It is not surprising that usgEn users did not show the same interest in a language that, perhaps, most of them did not readily comprehend.

OCR errors in the reo Māori extracted texts were listed as the most frustrating issue by the usgMi clients. Te reo Māori appeared to be an important issue for the usgMi clients. Navigation, searching and printing were mentioned more by usgEn than by usgMi clients.

In general usgMi clients seemed less likely than usgEn clients to offer negative comments about the digital library unless specifically asked, as in the fifth question. Where negative comments were made they were often made in English. UsgMi clients also appeared to offer more praise and note their appreciation for the Niupepa digital library, often using poetic phrases to do so.

7.3 Client Interview Methodology

As mentioned in the introduction to this chapter, two methods were used to obtain client feedback. This section describes the second method, client interviews. A full description is given of how the client interviews were conducted, followed by a description of how clients were selected to be interviewed.

7.3.1 Client Interview Procedure

Interviews were conducted with clients, or potential clients, of the Māori Niupepa digital library and the Hawaiian Nūpepa digital library. In total 16 interviews were conducted, involving between one and three participants per interview. Three interviews were with usgMi clients, five with usgHaw clients and eight with usgEn clients. A total of 30 people were interviewed.

The format, comprising three distinct stages, was consistent for all interviews. In the first stage the purpose of the research was explained; those being interviewed were then shown, and had explained to them, a Bill of Rights and a Research Consent Form, which they were asked to sign if they were happy to be interviewed. The Research Consent Form was also signed by the interviewer/researcher. A copy of the Bill of Rights and the Research Consent Form are listed in *Appendix C Interview Forms*.

The second stage of the interview involved an observational study. The participants were observed using the Māori Niupepa digital library or the Hawaiian Nūpepa digital library. Experienced clients were asked to demonstrate the ways in which they used the digital library. If participants had not used the Niupepa/Nūpepa digital library before, the purpose of the digital library was explained and then they were given some predefined tasks to complete. These tasks would require them to seek and retrieve information from the digital library. The interviewer generally remained in the background to observe and record, influencing the outcome as little as possible. At times the interviewer was required to explain the set tasks or to clarify why certain actions were taken. The observational stage of the interview lasted approximately 20 minutes.

The third and final stage of the interview involved asking the participants some prepared questions selected from the following:

- Why do you use the Niupepa collection?
- What types of information do you mostly seek?
- By which methods do you prefer to find information?
- By which methods do you prefer to browse information?
- Which language do you use when entering query terms?
- Do you alter the default language of the web site?
- Does the default language of the web site alter your language use?
- How often do you switch languages in a session?
- Do you have any issues, comments or thoughts about this digital library?
- Do you have any issues, comments or thoughts about indigenous language use of a bilingual web site?

Answering these questions took approximately 30 minutes. Participants were then given a small gift for taking part in the research.

Each interview stage was reviewed and the actions undertaken carefully documented. Items recorded included: the time spent on each particular activity; the success or otherwise of the activity; how the interface was used to access the information; what preferences were altered; what query operators and parameters were used; what difficulties were encountered, and how the indigenous language was used in the activities. Any comments or responses made during the observation and question/answer phases were carefully transcribed.

7.3.2 Finding Clients To Be Interviewed

In theory, identifying Māori Niupepa clients or Hawaiian Nūpepa clients for client interviews should have been difficult as no usage information was kept that distinguished precisely who the clients were. In practice, however, finding these clients was straightforward.

Many te reo Māori clients had provided feedback, both verbal and via email, to the Niupepa web site developers. These clients were contacted and agreed to participate in the client interviews. In fact, more te reo Māori clients were willing to be interviewed than time allowed.

Dr. Robert Stauffer, of Alu Like Inc.ⁱⁱ in Hawai‘i, was contacted and asked if he was aware of any Hawaiian Nūpepa digital library clients who would make suitable participants for this study. In a short time five interviews had been arranged with eight people who were known to Dr. Stauffer, all of whom used the Hawaiian Nūpepa digital library in the Hawaiian language.

Sourcing participants who used (or could use) the digital library in English was also not difficult. At the time participants were being sought, a lecturer of a

ii Alu Like Inc. was one of the organisations responsible for creating the Hawaiian Nūpepa digital library. A more detailed description of them is given in §6.1.2.

computer science undergraduate course circulated a request asking if his students could be interviewed in a usability study. A component of the students' course required them to experience the role of an interviewee. Consequently eight interviews were conducted with 19 computer science students, all of whom were asked to use the Māori Niupepa digital library. Although none of the students had used the digital library previously they were still able to offer some insightful comments. All had previous experience with using electronic resources and most had used electronic resources in a language other than English. Many of the students were bilingual or multilingual and most were visiting students from overseas, which meant their opinions arose from a wide range of perspectives.

Table 7-1 gives a summary of participants in the 16 interviews. The first three interviews were with usgMi clients who used the Māori Niupepa digital library in te reo Māori. The next five interviews were with usgHaw clients who used the Hawaiian Nūpepa digital library in Hawaiian. The final eight interviews were with computer science students who were introduced to the Māori Niupepa digital library and used it mostly in English. A description and summary of each interview is presented in the following section.

Summary of Users Interviewed				
	Digital Library	Usage Language	Participant	Nationalities
usgMi #1	Niuepepa	reo Māori	1 male	Aotearoa Māori
usgMi #2	Niuepepa	reo Māori	1 female	Aotearoa Māori
usgMi #3	Niuepepa	reo Māori	1 male	Aotearoa Māori
usgHaw #1	Nūpepa	Hawaiian	1 male	USA
usgHaw #2	Nūpepa	Hawaiian	1 female	USA
usgHaw #3	Nūpepa	Hawaiian	1 female 1 male	Hawai'i ⁱⁱⁱ
usgHaw #4	Nūpepa	Hawaiian	2 females	Hawai'i
usgHaw #5	Nūpepa	Hawaiian	1 female 1 male	England, Hawai'i
usgEn #1	Niuepepa	English	2 males	Aotearoa, South Africa
usgEn #2	Niuepepa	English	1 female 1 male	China
usgEn #3	Niuepepa	English	2 males	Aotearoa, Taiwan
usgEn #4	Niuepepa	English	2 males	China, Zimbabwe
usgEn #5	Niuepepa	English	2 males	Austria, Germany
usgEn #6	Niuepepa	English	3 males	Germany
usgEn #7	Niuepepa	English	1 female 2 males	Aotearoa
usgEn #8	Niuepepa	English	1 female 2 males	Austria, Aotearoa, Aotearoa Māori

Table 7-1: Summary information about the clients who were interviewed using the Niuepepa or Nūpepa digital libraries

iii While native Hawaiians could be classified as of USA nationality they are listed in this table as Hawai'i to distinguish indigenous Hawaiian people from non-indigenous residents of Hawaiian.

7.4 Client Interview Responses

This section gives a summary of each interview conducted with clients of the Māori Niupepa digital library and the Hawaiian Nūpepa digital library. The summary was taken from the interview documentation created when the video of each interview was analysed.

The setting of each interview and how the interview took place is explained. Some background information is given about each participant so that their responses can be considered in context. For example, a young participant who has a lot of experience using online electronic resources will give different responses from an older participant with little experience using online resources. Multilingual and bilingual participants are likely to respond differently to monolingual participants. Anonymity was maintained at all times during the interview process and in subsequent analysis; none of the interviewees are able to be identified from this summary.

This section is divided into four sub-sections. The interviews with usgMi clients are described first, followed by the interviews with the usgHaw clients and then with the usgEn computer science students. The final sub-section summarises the information detailed in the usgMi, usgHaw and usgEn interviews.

7.4.1 Interviews With UsgMi Clients

Three usgMi participants were individually interviewed as clients of the Niupepa digital library in te reo Māori. All were Māori; one was a native speaker of te reo Māori, the other two were second language learners and all spoke English fluently. Two were men and one was a woman and all had used the Niupepa digital library to various degrees. The interviews took place in Hamilton, Wellington and Christchurch (all cities in Aotearoa) at their places of work. Two interviews took place in October 2005 and the other in May 2006. The

participants remain anonymous. All interview recordings on all forms of media are scheduled to be destroyed by December 2007. Brief summaries of the three interviews follow.

UsgMi Interview #1

This participant was an academic in the field of te reo Māori, Māori customs and Māori history, in the over 50 years age bracket and with little experience using online collections. The participant had, however, used the Niupepa collection, perhaps six times a year for four years. The interview took 50 minutes in total, of which 29 minutes were usage observation and 21 minutes were spent responding to the prepared questions.

The participant was observed undertaking a topic query and examining six results in detail, with both the extracted text files and full size images being studied. The participant noted results on a piece of paper to be followed up at a later date and then submitted another query by entering the title of a periodical into the search box and pressing Rapu (Search). The results were examined, along with commentary information for the particular newspaper. Some time was spent trying to resize the font to make reading the texts easier. Some more documents were read until the conclusion of the observation session.

The prepared questions elicited the following responses: the Niupepa digital library was used to find suitable texts for translation exercises and to undertake tribal research; full text search was the method used to find most information because "it's easy"; reo Māori query terms were used as they matched the topic of the information being sought; initially the digital library was used in both English and te reo Māori but as "time went by I did not need to use that [switch to English] button"; the digital library was used in whatever language the interface default was set to.

UsgMi Interview #2

A tribal researcher working for a Māori tribal authority, in the 31–40 age bracket the participant was classified as a medium user of the Niupepa digital library, using it perhaps once a week for three years. Other online collections had also been used to the same extent. The observation part of the interview lasted for 18.5 minutes and responses took a further 10 minutes.

The participant gave query activity examples and entered a recent query for a particular name. The relevance of the results was discussed. Another query was undertaken, this time on a dialectal term, showing relevant results to the tribe for whom the participant was working. A third query was undertaken, for a historical letter, and then a fourth query on the participant's name; in both cases the results returned were relevant. The participant showed how the extracted text files were copied and used as material for the tribal web site and as sources for tribal educational material and traditional greetings. The participant also used the greetings in personal correspondence.

The participant then responded to the prepared questions: mostly the Niupepa digital library was used to find traditional tribal information that could then be used in a modern educational context, to increase personal reo Māori historical knowledge and to source examples of translations; full text search was the preferred method for finding information because "it's similar to Google"; the language interface was never altered but was always kept in te reo Māori "because I am looking for stuff in Māori"; only extracted text documents were requested as the participant was not interested in the original image; the participant had noticed errors in the extracted texts but was not aware that these were caused by the OCR process; the participant was not aware that quotation marks could be used to assist searching or of the English abstracts and discovered the commentary information only by chance.

UsgMi Interview #3

A research fellow, a history consultant and a translator, aged between 31 and 40 years, this participant was classed as a very heavy user of the Niupepa digital library, accessing it 3–4 hours per day, most working days for four years. The participant had similar experience with other online digital collections. The observation part of the interview lasted 40 minutes, with a further 16 minutes spent responding to the prepared questions.

The participant gave a comprehensive demonstration of using the Niupepa digital library, initially giving two examples of using the search engine to locate documents and altering the Some/All query parameter when necessary. Use of quotation marks to enable phrase searching was observed and examples were given of the Niupepa material being copied and pasted into word processing documents. Sourcing parallel language documents by concurrently entering query terms in te reo Māori and English was demonstrated. Specific documents were shown to be easily located when large blocks of text, copied from other documents, were entered into the query box. Further material about a subject was found by dividing the query term into component parts. For example, a query for ‘Whanganui’ was followed by a query for ‘Whanga nui’ which located more examples. The participant then demonstrated that the English abstracts sometimes contained information that was not located when undertaking a query solely on the newspaper documents.

The following responses were made to the prepared questions: the Niupepa digital library was used to find information on “all topics”, in particular, names, genealogy, traditional areas of residency and reports about what was said at gatherings; full text search was mostly used because “you get there quicker, to the actual particular topic that you are searching for”, however, browsing by date and by periodical listing was also undertaken; the commentary information and English abstracts were used; the participant mostly used reo Māori for the interface and queries, switching the interface language occasionally so that the reference material would be displayed in English and then could be easily copied and inserted into word processing

documents where references were required in English; on other bilingual web sites sometimes the interface language was switched to English "... I haven't got time to learn all those new words, it's laziness really ..."; however, the Niupepa digital library was predominantly used in te reo Māori as the terminology was readily understood.

7.4.2 Interviews With UsgHaw Clients

Five interviews were undertaken with eight users of the Hawaiian Nūpepa digital library. All participants used the digital library with the interface language set to Hawaiian, all had an understanding of written Hawaiian and all were residents in the islands of Hawai'i. Five participants were of Hawaiian ancestry. The interviews were conducted at their homes or places of work; one interview took place on the island of O'ahu and the other four were conducted on the island of Hawai'i. The interviews took place over a two day period in June 2006. All interviews were recorded onto digital video tape, enabling them to be fully documented at a later stage. The participants remain anonymous. All interview recordings on all forms of media are scheduled to be destroyed by December 2007. Brief summaries of the five interviews with usgHaw clients follow.

UsgHaw Interview #1

This participant was a male who did not have Hawaiian ancestry, who worked in the computer industry and was in the over 50 year age bracket. The Nūpepa digital library had been used to assist in his graduate and personal research. He was fluent in both spoken and written Hawaiian, used the Nūpepa digital library perhaps five times or more a week and other online resources perhaps two or three hours everyday. Regarding the interface language he stated, "The truth is, I don't know that I really pay attention to it, when I see the Hawaiian I don't change it, I suppose that if I saw it in English I am not sure that I would not change it either ... I

am happy to see it in Hawaiian.” The observation part of the interview lasted 25 minutes, with a further 20 minutes spent responding to the prepared questions.

The participant demonstrated using the Nūpepa collection to locate a citation by periodical name through the series list. He then clicked on the appropriate date and looked for an extracted text; as there was none he browsed the publication until he found the particular article that he was looking for. He then talked generally about some of the texts he had found in the newspapers and the wealth of literature available there. He demonstrated searching for certain topics and discussed the issues with phrase searching. He spoke about how he had browsed through some earlier newspapers and discovered some of the earliest examples of indigenous language publications by native speakers. He then discussed issues regarding accessing later editions of newspapers.

The following responses were made to the prepared questions: the Nūpepa collection was used to access articles for which he had the citation and to locate new information; the default language of the digital library wouldn't affect usage, “... I am goal driven when I get there ...” but that if the site was suddenly defaulted to the English language he “...would possibly switch back from English to Hawaiian because I suppose that is what I am used to...”; while aware of the purpose of the Some/All query parameter he didn't use it much as he could not “... say that it makes clear sense to me that my searches show up a lot differently for me ...”; the participant would have used the quotation marks (phrase defining query operator) had they been working.

UsgHaw Interview #2

This participant was an information provider, in the over 50 year age bracket and very competent with computers. Of non-Hawaiian ancestry, she had a minimal spoken fluency and medium written fluency in the Hawaiian language. The participant was a heavy user of the Nūpepa collection using it 1–2 hours per day, most days and used other online collections to a similar extent. The observation part of the interview lasted 22 minutes, with a further 15 minutes spent responding to the prepared questions.

The participant was observed undertaking a regular transcription exercise; a Nūpepa document was located by the series list and a transcription of the article was created in a Microsoft Word file. The participant explained how references were obtained from other online and physical sources and then the document was viewed in the Nūpepa web site. The series list was displayed and the publications for each periodical listed. The participant then selected the appropriate time period, requested an article and then began reading the article to locate the topic of interest.

The following responses were made to the prepared questions: the query facility was not used as the participant knew how few of the newspaper documents had been transcribed; query parameters were not altered as the participant was unaware of them and of the detach link^{iv}; usage of the digital library was in Hawaiian because "... I speak about 6 or 7 languages fluently and to switch back and fourth is time consuming. If you can work in the same language you don't have to waste brain time translating".

^{iv} A link that would open the current document view into a new browser window.

UsgHaw Interview #3

The participants were a Hawaiian male and a Hawaiian female. One was a professor in the 31–40 age bracket, the other an editor/researcher/translator in the 21–30 age bracket. Both were fluent in written and spoken Hawaiian. Each accessed the Nūpepa digital library about five times a week and other online resources perhaps 6–12 times per day. The default language of the digital library was always set to Hawaiian although one participant altered this setting to English on the very first visit in order to understand what one particular menu item was referring to. The observation part of the interview lasted 28 minutes, with a further 22 minutes spent responding to the prepared questions.

The participants demonstrated submitting a query for a particular term so they could examine the structure of sentences in which the term was used. They then gave the example of how they had used the Nūpepa collection to find information about an ancestor and stressed how relevant the results had been. One participant explained how students had been required to find articles from the Nūpepa digital library and then to transcribe these into modern Hawaiian. The participants demonstrated the amount of information that could be obtained from transcripts of songs and chants. The participants also showed how they undertook browsing, both on previously referenced articles and on newspapers where particular content or a writer/editor was known.

The following responses were made to the prepared questions: both participants had similar strategies and reasons for accessing the Nūpepa digital library; browsing was preferred when the periodical name and date was known; querying was preferred when it was not known where, or even if, the information existed in the Nūpepa collection "... but then you have the problem that not all the pages are searchable. Am still waiting for that"; neither participants would alter the default language of the digital library, "... I don't see it as being in one language or the other ... I don't really pay attention to the things that I don't need to use ..."; both participants had tried changing the Some/All setting on the query page

but it did not seem to make a difference; quotation marks were used to enable phrase searching, but one participant felt that this did not always seem to work.

UsgHaw Interview #4

Two Hawaiian females participants in this interview, one under 21 years old and one aged over 50 years. One was involved with teacher training, the other a tertiary student and part time writer. Both were fluent in written and spoken Hawaiian. The student used the Nūpepa digital library perhaps 3–5 times a week, the teacher trainer perhaps only twice during weekends. Neither was aware that the language of the interface could be set to anything but Hawaiian or that non-Hawaiian texts (i.e. English) were available in the Nūpepa collection. The observation part of the interview lasted 20 minutes, with a further 24 minutes spent responding to the prepared questions.

The observation section of the interview consisted more of discussion with the interviewer than strict observation of usage, although, the participants did show how they accessed documents. One used other online resources to find citations to the newspapers and then used the Nūpepa series browse and a date identification to select the newspaper required. She then either read the document on the screen or printed it out in a large size so that she could take time to read it. The other mentioned that the image format of the documents was “nice” but as she only used the search facility she mostly accessed the extracted text documents; “... it’s way easier to read and the search terms are highlighted.” Examples were shown of both access methods.

Both participants were surprised that the interface and some articles were available in English. Neither was aware of the query parameters (Some/All) and of the option to detach an article. Both participants stressed how important this resource was for the Hawaiian people.

UsgHaw Interview #5

These two participants were a male (with no Hawaiian ancestry) living in Hawai‘i and a Hawaiian female,^v both in the 31–40 age bracket and fluent in spoken and written Hawaiian. Both participants worked with curriculum development and one was also involved with translation, editing and technical support. One participant was a medium to heavy user of the Nūpepa digital library, using it only once or twice per week but using it for three or four hours at a time and the other participant had not used the Nūpepa collection but had seen it being used. Each was considered a heavy user of other online collections and resources, using them up to 20 times a day. The observation part of the interview lasted 23 minutes, with a further 16 minutes spent responding to the prepared questions.

The Nūpepa digital library was demonstrated primarily by the participant who had used it frequently. Examples were shown of using the digital library to find patterns of word usage and looking up uncommon terms. The participant explained that phrase searching with quotation marks was really useful but did not appear to be working much of the time and mentioned the appropriateness of the icons and accessing extracted text and facsimiles. The participant regularly accessed full size images and used the zoom feature to decipher exactly what some words were or printed out and enlarged the images to assist with deciphering what some characters and words were.

The following responses were made to the prepared questions: these participants were an example of usgHaw clients who preferred to submit queries to find documents but understood that the search facility only searched on a small percentage of the documents and stated “I wish this would get bigger faster”; the Some/All query parameter was not altered, rather the preference was to see all of the query terms all of the time; full size images were preferred despite spending “... twice to three times as long looking at the image ... it’s harder to read but I think it is more accurate. That’s why I do it ... Sometimes I can get led astray by

^v Hawaiian female and Hawaiian male are terms used in this chapter to designate ethnicity not country of residence.

seeing the printed word [OCR extracted text format] and thinking it is accurate"; some of the links were unclear so were not used (for example, mid-size image icon, detach icon) " ...I don't know what they mean so I never use them."

7.4.3 Interviews With UsgEn Clients

Students undertaking computer science degrees at the University of Waikato were interviewed to provide user feedback on the Niupepa digital library from non-indigenous and international perspectives. This particular group was selected because of their availability and the wide variety of nationalities they represented. Although none spoke or read te reo Māori, many of the students were bilingual or multilingual and, consequently were able to provide feedback on their personal experience using electronic resources in a language other than English.

Eight interviews with 19 participants were held over a four day period in March 2006; each interview involved either two or three participants. Although none of the students had used the Niupepa digital library, all had considerable knowledge and experience using online databases. The interviews were conducted in a usability lab where the students' physical actions, verbal interactions and online activity were all videoed simultaneously.

As none of the students had used the Niupepa digital library, the interview format was adapted to include an introduction to the digital library, its purpose, what types of information were available and how they could be accessed. Time was given to the participants to familiarise themselves with the digital library.

The participants remain anonymous. All interview recordings on all forms of media are scheduled to be destroyed by December 2007. Brief summaries of these eight usgEn interviews follow.

UsgEn Interview #1

Both participants were male, aged between 21 and 30 years old. One was from South Africa and the other from Aotearoa. Both were heavy users of online collections, using them about 30 times per week. The interview lasted for 45 minutes; 35 minutes of using the digital library and 10 minutes of answering questions about usage.

After the familiarisation period the participants were asked to locate documents relating to specified topics: documents relating to ‘measles’, to some of their ancestors and to a place called ‘Te Awamutu’. The participants were quick to use the Some/All query parameters and also made use of the quotation marks to undertake phrase querying.

In the final part of the interview the participants were asked questions about their (brief) usage of the Niupepa digital library. They were unclear about the concept and purpose of an English abstract and selected this option for searching as they mistakenly thought a query would return only documents in English. They were not sure how the English abstract texts differed from the texts of the newspaper documents. As these participants could only speak English they were only interested in documents written in English. They suggested that the addition of an icon showing the language of the document in the results and series list pages.

UsgEn Interview #2

One female and one male were interviewed. They were both from China, with Mandarin as their native language and were in the 21–30 age bracket. They were heavy users of online collections, using them about 20–30 times per week. The interview lasted for 47 minutes; 37 minutes of using the digital library and 10 minutes of answering questions about usage.

After the introduction they spent about 10 minutes exploring the interface, reading the home page, the help page and the preferences page. The interface language was switched to Mandarin and a query submitted for the term “Waikato”. The results were examined but they were unsure what some of the numbering in the results pages referred to; when the interface language was switched back to English the numbering scheme became clearer. The participants were asked to find information on measles, a disaster and a famous person. These queries were all undertaken competently.

These participants had some slightly different approaches from those observed in other interviews. Before the activity began they took time to read the about, help and preferences pages where they discovered the option of having the site in Mandarin. The participants remarked that when they were in their own country, China, they preferred to undertake searching activity in Mandarin. Part of the reason they came to Aotearoa was to improve their English and, because assignments had to be written in English, they preferred to undertake this task in English. They also seemed to click through the content very quickly, not pausing very long to read and browse the content of the Niupepa documents.

UsgEn Interview #3

The participants were two male students in the 21–30 age bracket, one from Aotearoa (non-Māori) and the other from Taiwan. Neither was literate in te reo Māori; the Taiwanese participant also spoke Mandarin. Both were heavy users of online collections and online resources, stating they used these over 100 times per week. The interview lasted 35 minutes; 28 minutes of using the digital library and 7 minutes of answering questions about usage.

After the initial introduction, they spent 11 minutes looking around the site and trying out many of the interface controls and browsing options. A very short time was spent looking at each of the interface pages and even less time looking at newspaper documents not written in English. The participants were asked to find three separate pieces of information on diseases, three on natural disasters and

some information on an ancestor, all of which were completed within 14 minutes. These tasks were then followed by the question and answer section of the interview.

Some important points were raised in the interview. These participants preferred to search on the English abstracts as these were “more convenient, if we looked at these we had a better understanding about what it was going on about.” They suggested that in addition to a query term being highlighted in the results page, the page should be automatically scrolled so the query term becomes immediately visible to the user. The bilingual participant mentioned that he only uses a Mandarin interface when he knows that the articles that he is searching for are in Mandarin.

UsgEn Interview #4

The participants were two male students in the 21–30 year old age bracket, one from Zimbabwe who also spoke Afrikaans and the other from China who also spoke Mandarin. Both participants were heavy users of online collections and online resources estimating their use at over 50 times per week. The interview lasted 45 minutes; 35 minutes of using the digital library and 10 minutes of answering questions about usage.

After the introduction they spent 14 minutes exploring the interface, the different browsing options available, submitted some queries and examined the results. They were then asked to find three separate pieces of information about diseases. This task took longer than in some of the other interviews as they spent a lot of time reading the documents. They were then asked to search for information on three disasters and found documents about an earthquake, a flood and a tidal wave. The Mandarin version of the interface was tried and an incorrect translation for volume number was noticed. This concluded the observation section of the interview.

The participants raised some new issues about the usability of the Niupepa digital library. One complained about the (green) colour of the interface and felt that if he had to look at it for too long his eyes would become tired. There was uncertainty whether some of the buttons were clickable. For consistency with other web sites a site logo should be added, the help button should be located on the far right and the home page button on the far left. The Chinese participant explained that usually he uses the Web in Mandarin as it is easier for him to understand. However, if he wanted to search for a sport, such as rugby, then he "cannot search the Chinese [language] website[s] because many Chinese people do not like rugby."

UsgEn Interview #5

The participants were two 21–30 year old male students, one from Germany and the other from Austria. Both also spoke German and were heavy users of online collections and online resources, estimating their use at over 100 times per week. The interview lasted 45 minutes; 32 minutes of using the digital library and 13 minutes of answering questions about usage.

After the introduction they were asked to familiarise themselves with the Niupepa interface. First they set the interface language to German and looked at the different methods for accessing documents and examined some of the documents. They submitted some queries and examined the results. The participants were then asked to find three separate articles relating to diseases. Their initial spelling was poor so they used a German-English online dictionary to get correct spelling, then using the Niupepa query facility found three articles about diseases. The next tasks were to find information on natural disasters and an article about a person. They located articles about three separate earthquakes and about James Cook. Finding these articles completed the observation section.

Some interesting behaviours and responses were recorded in this interview. It was the first where the interface was used for the entire interview in a foreign (non-English, non-Māori) language. The participants only searched the English

abstracts, perhaps because they did not want the results page to be cluttered up with articles that they could not read. They noted that the English abstract page numbering did not always link correctly to the associated newspaper page. Both participants preferred using German when browsing/searching in their free time and English when undertaking course-related work, presumably because their course was taught in English.

UsgEn Interview #6

The participants were three 21–30 year old male German students with English as their second language. All participants described themselves as heavy users of online resources, each accessing these at least 50 times per week. The interview lasted 47 minutes; 32 minutes of using the digital library and 15 minutes of answering questions about usage.

After the introduction they spent 15 minutes familiarising themselves with the interface. The preferences were examined and the interface language was initially switched to German but then switched back to English. They looked at the different browsing methods, accessed several documents, looked at some of the query parameters, submitted a query and then examined some of the query results. The participants were then asked to find articles on three separate illnesses. They entered ‘illness’ into the search engine and searched only on English abstracts. They were surprised that the query term was not found until the interviewer explained that English abstracts only searched on reo Māori documents. The query scope preference was adjusted to ‘papers, commentaries and abstracts’ and then articles were located. Three articles about disasters and an article about a famous person were located without difficulty.

The participants displayed some pertinent usage behaviours. After trying the interface in German they quickly decided to switch it back to English, because they noticed that many sections of the help page had not been translated from English into German. “That’s [in English] how we are used to interact with the computer anyway ... most of the time”. When they first

accessed the series page everything appeared to be in te reo Māori, which was initially daunting and the comment "Let's go back" was made in a joking fashion. The participants made comments about the usability of the interface, the placement of buttons, the labels and what could be added to make the site more usable. They also commented that they rarely undertook language switching of an online database, preferring to use the database in what appears to be the language in which it was built. They also mentioned that if a web site was not easily understandable then they would quickly look for another site.

UsgEn Interview #7

The participants were three 21–30 year old students, one female and two males. They were all from Aotearoa, although one had spent a considerable part of his life in the United States. None was of Māori descent and all described themselves as heavy users of online resources, each accessing online resources between 50 and 100 times per week. The interview lasted 40 minutes; 27 minutes of using the digital library and 13 minutes of answering questions about usage.

After the introduction they were given time to explore the site and looked at the different methods for accessing the documents. Using the date option, they accessed a document and spent time reading the text. They took time to read the help and preferences pages. As one participant had Japanese knowledge they briefly looked at the interface in Japanese before switching it back to English. They returned to the Series option, accessed a publication, looked at the English abstract, and read some background information about the English Abstracts. The participants were then asked to find information about three diseases which they did, although they did have some difficulty trying to spell the diseases correctly. Next they were asked to find information on three natural disasters, which they did without difficulty using the query facility. The final task was to search for a surname which was undertaken without difficulty.

These participants exhibited some unique behaviour. More time was spent reading the Niupepa documents perhaps because the participants were from Aotearoa the

documents had more relevance to them. The series list did not seem to appear so foreign, perhaps because many of the reo Māori titles were familiar to these students. The participants were not particularly interested in the full size images "... if the text was accurate then I wouldn't have any reason to view the full size image ...". They found that they were unable to enter wildcard characters with query terms when undertaking searching. They suggested a spell checker on the search interface would be helpful. There was also the suggestion that pages with highlighted query terms should be automatically scrolled so that the terms were instantly viewable, "generally that's why you go to search pages to go straight to that text".

UsgEn Interview #8

The participants were three students, one female and two males. The female was Austrian and fluent in German and English. She was in the 21–30 age bracket and used online collections or searching perhaps 500 times a week. One of the males was from Aotearoa (non-Māori) in the 41–50 age bracket. He was a part time student, part time stores man who used online resources perhaps 150 times a week. He had no fluency in te reo Māori. The other participant was an Aotearoa Māori with tribal affiliations to Ngāti Porou and Ngāti Raukawa. He had a basic understanding of both spoken and written reo Māori. An educationalist, he accessed online resources perhaps 50 times per week. The interview lasted for 48 minutes; 29 minutes of using the digital library and 19 minutes of answering questions about usage.

After the introduction they were asked to explore the site. They began by looking for information about the Tarawera eruption that occurred in Aotearoa in 1886. Some initial queries returned results but without a preview of the query results the participants were unsure whether the documents were relevant and did not access them. A series browse at first appeared daunting but the titles of the newspapers were explained by the Māori participant and some documents were examined. Two students were happy to see text but one wanted to see the actual image. The

participants undertook some searching on the English abstracts and tried altering the Some/All option. They were asked to find three diseases and to search for some family names, both of which were done without difficulty. In one instance quotation marks had to be used to narrow the results. The family name search concluded the observation portion of the interview.

Some interesting behaviours were observed. When looking at the series list one participant stated, "Problem is none of these mean anything to me, all those results". The Māori participant replied "...this is just giving the title of the newspaper and the volume". Clearly, having some familiarity with te reo Māori improved the usability of this interface in English. It was clear that the participants from Aotearoa were more interested in the texts and wanted to spend longer reading the content. They gave feedback on improving the usability of the interface, for example clearer browsing buttons, a different layout of English abstracts and clearer navigation.

7.4.4 Client Interviews Summary

Clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries were interviewed using two methods. First, the clients were observed using the digital library. Second, the clients were asked questions about why and how they used the digital library. Observations and responses were analysed to determine what factors influenced the clients' use of a particular language.

Three client types were interviewed: those who used the Māori Niupepa digital library with the interface language set to reo Māori (usgMi) all of whom were Māori; those who used the Hawaiian Nūpepa digital library with the interface language set to Hawaiian (usgHaw), both Hawaiian and non-Hawaiian clients; and those who used the Māori Niupepa digital library with the interface language set to English (usgEn), both Māori and non-Māori clients.

The usgMi and usgHaw clients interviewed were experienced users of their particular digital library, all were bilingual and understood the interface in both

the indigenous and non-indigenous languages and were able to read the collection documents in both languages. The usgEn clients interviewed were all experienced users of online electronic resources but none had previous experience with the Māori Niupepa digital library and, aside from one, none were literate in the indigenous (reo Māori) language which meant that they could not use the interface in te reo Māori or comprehend any reo Māori content of the digital library.

With the differences in client types in mind, the factors which affected the use of a particular language in a bilingual digital library environment are summarised in the next three sub-sections.

UsgMi Client Interviews

Only three usgMi clients were interviewed, insufficient to represent all clients who used the Niupepa digital library in te reo Māori. However, these interviews have provided some evidence to suggest factors affecting te reo Māori usage by clients of the Māori Niupepa digital library.

The usgMi participants demonstrated how the Niupepa digital library was used for tribal, personal and academic research. Use of the query facility method was most commonly chosen by the two younger participants, whereas the older participant had a stronger preference for accessing the documents using the series list. One of the younger usgMi participants only accessed the extracted text versions of documents whereas the older usgMi participant mostly accessed the full size images.

The usgMi participant most experienced with the Niupepa digital library, and other online electronic resources, made the most use of query parameters and query operators also using novel ways to improve his querying results, such as using query terms in two languages, separating a single query term into two portions and submitting queries on blocks of texts. The other two usgMi participants made little use of query parameters and query operators.

All usgMi participants showed great interest in the documents located and, consequently, a lot of time was spent reading the documents. They did not mention any negative issues with the Niupepa interface (in te reo Māori), despite two of the participants clearly demonstrating difficulties using the Māori Niupepa interface. One participant had problems with viewing a full size image on the screen, another undertaking queries for phrases. However, neither raised these as problems until they were later queried further on these issues.

All three participants used the interface in te reo Māori because they were searching for Māori topics and felt comfortable in that language. Two mentioned that they used the English version of a bilingual interface if the reo Māori version was not clear.

UsgHaw Client Interviews

Eight usgHaw clients were interviewed. Again, this number is insufficient to represent all clients who used the Nūpepa digital library in Hawaiian, but has provided some evidence suggesting factors affecting Hawaiian language usage by clients of the Hawaiian Nūpepa digital library.

The usgHaw participants demonstrated how the Hawaiian Nūpepa digital library was used for personal research, for education purposes, for Hawaiian language studies and as a source of material for publications. The Nūpepa documents were accessed from browsing by series and by submitting a query. The usgHaw participants generally showed an awareness that only a small proportion of documents had been converted to text and so were prepared to use means other than the query facility to locate documents. In many instances they knew a particular article existed, through references or having previously accessed the article. Consequently, the series list was frequently used to locate an article.

Those with more experience with the digital library tended to view the midsize and full size formats. Those who undertook queries tended to view the extracted

text formats. Querying and use of extracted texts was favoured by the younger participants.

Experienced participants had tried to use the query parameters and query operators but found that they either did not work or did not work as expected and were not used subsequently. Documents relevant or otherwise interesting were often saved to their local computer and/or printed out for more intensive study.

The usgHaw participants did not mention any negative issues with the Hawaiian Nūpepa interface, aside from a common wish to see more documents with extracted text files (thus allowing them to be found by a query). Many stressed how important this resource was for the Hawaiian language and for research into Hawaiian history and culture.

All eight participants used the interface in the Hawaiian language for several reasons; Hawaiian was the default language setting of the Nūpepa interface, as the documents were (virtually all) written in Hawaiian it was cognitively less challenging, and they all expressed a desire to support the use of the Hawaiian language. Some of the participants mentioned (and demonstrated) that they had briefly switched the language of the interface when some terminology used in the Hawaiian language interface was not understood.

UsgEn Client Interviews

Nineteen people were interviewed as clients who used the Māori Niupepa digital library in English. While classified as usgEn clients, they were not representative of all usgEn clients as they were first time users and undertaking set tasks rather than using the digital library through choice. Despite these differences, the usgEn clients interviewed were still able to provide some opinions on using the Niupepa digital library in English and offer some thoughts on the non-English use of online electronic resources.

None of the usgEn participants used the Māori Niupepa digital library in Māori as all but one was unable to understand te reo Māori. The digital library was tried in other languages; one group used it in German for the whole of the interview and other interview groups tried the interface in Mandarin (twice), German and Japanese before switching it back to English.

The usgEn participants were instructed to find information based on certain topics and all used the query facility to achieve this. The series list displayed all the periodical titles but as almost all of these titles were written in te reo Māori (see *Figure 3-4*) this was quite foreign to the usgEn participants. On seeing the series list most chose to return to a page where most of the text was written in English.

The usgEn participants, all experienced users of online electronic resources, readily used the Some/All query parameter and the phrase defining query operator. Many wondered if a spell checker could also be incorporated into the interface. Those who were not from Aotearoa spent little time reading the Māori Niupepa documents while those from Aotearoa often noticed articles of interest to them and subsequently spent a longer time reading documents of the Niupepa digital library.

Many issues about the digital library interface were raised by the usgEn participants: the concept and purpose of the English abstracts was unclear to many of them; they mentioned the query activity would be improved if preview documents were shown in the results page; suggestions were made to improve many aspects of the interface, including where to use more icons, the placement of buttons and links, the colour of the interface and clearer navigation. It seemed the usgEn clients were less interested in the content of the digital library and more interested in critiquing the usability of the interface.

The multilingual usgEn clients mentioned factors which led to their decision to use an electronic resource in a particular language. These included matching the interface language to the content of the resource, matching the interface language to the predominate language of the interface, a desire to use the interface in one particular language (for example, to match the language that the student's course was taught in) and using an interface language that was easily understood.

7.5 Summary

This chapter has discussed two methods used to generate feedback from clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries. Sixty responses were received to a web questionnaire placed on the Niupepa web site in 2006. Sixteen observational studies and interviews were undertaken with clients who used the digital libraries. The web questionnaire responses were too low in number and the clients interviewed too few for the feedback to be representative of all Niupepa/Nūpepa clients. However, the feedback generated has been valuable for two reasons. First, it raised some important usability issues with indigenous languages on bilingual digital libraries. These usability issues are summarised in Chapter 8. Second, the feedback has suggested reasons why indigenous languages are used by clients of digital libraries; specifically why indigenous language speaking clients used these particular digital libraries and why these clients chose to use these digital libraries in the indigenous language.

Feedback from the web questionnaire and client interviews suggests that the Māori Niupepa digital library was used by te reo Māori clients primarily for undertaking tribal research, research on te reo Māori and/or on personal or family information. Research on historic practices, assistance with translations and recreational reading were also reported as reasons that this digital library was used by te reo Māori clients. The range of industries accessing the Māori Niupepa digital library appeared to be mostly limited to academic institutions and tribal organisations with personal researchers also using the resource. The Māori Niupepa digital library appeared to be used in te reo Māori by a small range of industries for a wide range of topics.

Feedback from the client interviews suggests that the Hawaiian Nūpepa digital library has been used extensively in the Hawaiian language education institutions, both at secondary and tertiary levels, often as a prescribed teaching resource, for studies on Hawaiian language, history and culture. In addition the Hawaiian Nūpepa digital library has been used regularly as a source of written material for Hawaiian language publications. The Hawaiian Nūpepa digital library appeared to

be used in the Hawaiian language by a small number of industries for a wide range of topics.

Indigenous language clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries have emphasised how important the digital libraries are because of the content they contain. Clients reported using the content to assist with discussions on history, cultural practices, land tenure and genealogy. Some of the content had even been presented in courts of law. The second aspect was the ‘treasure of the language’ available on the newspaper pages. Indigenous language clients lauded the quality and simplicity of the written indigenous language writings and suggested that making this available assisted in the revival of the traditional indigenous language. Some comments by indigenous language clients include:

“A great source of written information pertaining to the lives of my tupuna [ancestors], iwi [tribe], hapu [sub-tribe] and whanau [family]. In addition, the papers themselves provide interesting accounts of the social and political context of Maori communities at that time. This is incredibly valuable.”

“Kare i tua atu hei patengi kai ma Hinengaro. Ko te mihi ki a Nehe ma i waiho mai e ratou enei tu korero hei whakamaro i toku reo ake kia rite ki ta ratou i korero ai ...”

[There is no greater database to nourish the mind. Salutations to those of the past who left behind these words to strengthen my own (indigenous) language so that it can become similar to the language spoken by them...]

Thus, the content attracted many indigenous language clients to use these two digital libraries. Why did these clients choose to use these electronic resources in the indigenous language?

Feedback from usgMi and usgHaw clients suggested that the indigenous language was used for three primary reasons:

- the content was mostly in the indigenous language and it was cognitively less challenging to have the interface also in the indigenous language
- as the default language of the interface was normally set to the indigenous language, clients using the digital library would often not notice which interface language it was set to
- clients wanted to support the use of the indigenous language. They stated they want to use the indigenous language in as many avenues as possible and digital libraries provide another avenue to do this.

Indigenous language clients also mentioned that if the indigenous language version of the interface was not clear they would switch to the English language version to understand what was being displayed. Once the interface was understood it was then generally switched back to the indigenous language for the reasons listed above.

Feedback from the multilingual usgEn participants also supported this view, suggesting that they also used an interface in a particular language because of the content, the default language setting and the desire to use one particular language. This feedback suggests that content, default language setting and desire to support indigenous language use affected all multilingual clients (both indigenous and non-indigenous) when making decisions about which language to use in a multilingual online interface.

This concludes the chapter about why indigenous languages were used by clients of digital libraries. The final chapter summarises the findings of this thesis and the thesis statement will be revisited.

8 Research Findings

The research described in this thesis has examined indigenous language usage in a digital library environment that has been accessed via the Internet.

Using the Māori Niupepa and Hawaiian Nūpepa digital libraries, this research has shown how indigenous languages were used by choice in these electronic environments during 2005. The results provide encouragement and optimism to people who are striving to retain, revitalise and increase the use of indigenous languages in information technologies. The Transaction Log Analysis (TLA) methods used in this research serve as an example of how web logs can be used to provide significant information about language usage in bilingual online information systems. Combining the TLA with user feedback has provided insights into how and why clients use indigenous languages in their information retrieval activities. These insights in turn, show good practise that is relevant not only to those working with indigenous languages, indigenous peoples or multilingual environments, but to all information technology designers who strive for universal usability.

This thesis comprised eight chapters. The first chapter described why it is important to investigate the use of indigenous languages in electronic environments and proposed that digital libraries can assist in the usage, and subsequent revitalisation, of these languages. The second chapter introduced TLA and explained how this methodology was used to generate data about language usage. This methodology was then used in the next four chapters to analyse: te reo Māori usage in the Niupepa digital library (Chapter 3); international usage of te reo Māori in the Niupepa digital library (Chapter 4); the effect of the default

language setting on te reo Māori usage (Chapter 5) and on Hawaiian language usage of the Nūpepa digital library (Chapter 6). Some user feedback was described in Chapter 7, which suggested reasons why clients used the indigenous language in digital libraries. Finally, Chapter 8 summarises the key findings from the previous chapters and returns to the thesis claim that digital libraries can be used to assist in the usage of indigenous languages.

8.1 Reo Māori Usage

Te reo Māori was used in the digital library environment.

Results from TLA show that te reo Māori was used in 37,936 requests to the Niupepa digital library in 2005 an average of over 100 requests in te reo Māori per day. Te reo Māori requests to the Niupepa digital library comprised 20.5% of the total requests.

It is difficult to accurately calculate the percentage of potential Internet clients who have an ability to undertake online activities in te reo Māori. A 2001 study indicated that only 35% of Māori (in Aotearoa) had access to the Internet compared with 51% of New Zealand Europeans (Te Puni Kōkiri, 2001a). The 2006 census indicated that only 24.4% of Māori were literate in te reo Māori (Statistics New Zealand, 2006). Applying, with some reservations, the 2001 access figures to the 2006 census numbers:ⁱ

- 35% * 565,329 = 197,865 Māori with Internet access
- 24.4% * 197,865 = 48,279 Māori speaking people with Internet access
- 65% * 3,462,618 = 2,250,701 non-Māori people with Internet Access
- 2,298,980 of population in Aotearoa have Internet Access
- 2% of the Aotearoa Internet Access population speak te reo Māori.

ⁱ These figures must be treated with caution. No allowance has been made for the increase in Internet users from 2001 to 2006, for the predominant age group of Māori speakers and online users, for the type of Internet access, or for the non-Māori speakers of te reo Māori.

The conservative estimate calculated above suggests that as little as two percent of the population in Aotearoa have the ability to access and also understand online information in te reo Māori. Given this low percentage, and without considering other factors such as who would want to access the Niupepa, the fact that over 20% of Niupepa requests were with a reo Māori language setting suggests that this digital library supports, and perhaps, encourages, usage in te reo Māori.

User feedback has suggested that the content of the web site, the default interface language and the desire to use the web site in te reo Māori have been factors which led to use in te reo Māori.

8.2 International Reo Māori Usage

Te reo Māori was used by non-Aotearoa clients.

Results from TLA showed that approximately 10% of requests to the Niupepa digital library were requests from international (non-Aotearoa) clients. International clients were almost three times more likely to use the interface in te reo Māori than national clients. Of the 17,390 international requests, 45.4% were with the interface language set to reo Māori while of the 155,515 national requests, 16.5% had the interface language set to reo Māori. International clients were most likely to use the interface in te reo Māori when the default language of the web site was set to reo Māori.

International clients who used the Niupepa digital library in te reo Māori exhibited different behaviours from national clients who used te reo Māori. International clients appeared to be experienced in information retrieval and confident to use the interface in English or te reo Māori. The use of query parameters and query operators and the apparent accuracy of the query activities indicated that these clients were familiar with information retrieval activities. A comparatively higher proportion of requests for parallel (reo Māori/English) documents and English language documents suggested that international clients were seeking information for English language purposes more often than national clients.

Use of te reo Māori by international clients in the Niupepa digital library is an example of how an electronic resource such as the Niupepa digital library can be used to assist in the international usage of an indigenous language.

8.3 Default Language Effect

Setting the default language of the interface to an indigenous language increased clients' use of the indigenous language.

Analysis has shown that there were 21% more reo Māori usage requests when the default language was set to te reo Māori. These requests were made by approximately the same number of clients, indicating that setting the default language to reo Māori did not encourage new clients to use te reo Māori, rather it encouraged bilingual clients to use te reo Māori more often.

When the default language was set to te reo Māori the reo Māori usage sessions were longer with more requests. These sessions also contained more browsing by series and more requests for the extracted text formats of documents. In addition, when the default language was set to te reo Māori, clients classified as reo Māori clients requested significantly more documents written in te reo Māori and significantly fewer documents written in parallel text (reo Māori and English). These clients also submitted more queries, more query terms per query and used a higher proportion of te reo Māori query terms.

These results indicate that setting the default language of a bilingual interface to an indigenous language not only results in an increase in use of the indigenous language but also promotes more situations where the indigenous language can be used. Clients commented that it was good that the indigenous language was the default language; they quickly became accustomed to it and would switch back to it if they found the indigenous language was not the default language.

8.4 Hawaiian Language Usage

The Hawaiian language was used in the digital library environment.

Results from TLA undertaken on the Hawaiian Nūpepa digital library showed that the Hawaiian language was used in 112,786 requests in 2005, 53.4% of the total (filtered) Hawaiian Nūpepa requests. When the requests were classified into language of usage categories, approximately half were Hawaiian usage requests, one quarter were English usage requests and the remaining quarter were classified as bilingual usage requests.

Feedback from Hawaiian language users, similar to the reo Māori feedback, has suggested that this digital library was used in the Hawaiian language because it was the default language of the interface, it matched the content of the digital library and the users expressed a desire to use the Hawaiian language in as many avenues as possible.

While many of the information retrieval behaviours of Hawaiian language clients matched those of reo Māori clients there were some notable differences. There were more indigenous language requests to the Hawaiian Nūpepa digital library than to the Māori Niupepa digital library. One reason for this difference could be attributed to the content of the respective digital libraries. The Hawaiian Niupepa digital library content was primarily in the Hawaiian language with an insignificant amount available that was written in English. Consequently, this digital library has little value to clients who are not literate in the Hawaiian language. The Māori Niupepa digital library has 46% of its content written in parallel text (reo Māori and English) and a further 23% contains English abstracts explaining what the reo Māori content is about. Consequently, over two thirds of the Niupepa content has value to clients who are not literate in te reo Māori. As many more potential clients who were monolingual (English literate) would be able to understand much of the content it seems likely that this larger group of clients would submit the most number of requests.

8.5 Indigenous and non-Indigenous Usage

Indigenous language clients exhibited different information retrieval behaviours from non-indigenous language clients

TLA revealed that the information retrieval preferences of indigenous language clients were different from the preferences of non-indigenous language clients. While both types of clients predominantly used queries to locate documents, the non-indigenous language clients had a higher preference for this method. The indigenous language clients had a stronger preference for locating documents by the series list. User feedback suggested that this was because the indigenous language clients often had prior knowledge of which publication series they were interested in. As the publications were mostly indigenous language titles, the series list was mostly written in the indigenous language, which tended to alienate the non-indigenous language clients. The non-indigenous language clients also had a higher propensity for accessing documents by using the date facility, whereas the indigenous language clients tended not to use this method. This difference could be attributed to different values and different attitudes that cultures display towards time (Hall, 1983).

Indigenous language clients preferred to access documents in the indigenous language while non-indigenous (English) language clients preferred to access parallel language documents and English language documents. Statistics from TLA highlighted this behaviour and it was not unexpected.

Queries undertaken by indigenous language clients often involved a greater number of query terms. It can be expected that clients are likely to enter more query terms in a language they are familiar with. A higher number of query terms entered by indigenous language clients meant a correspondingly greater use of the Some/All query parameter and of the quotation marks phrase matching operator. Query terms submitted by indigenous language clients were mostly in the

indigenous language, matching the language of the content, and appeared to be more successful than non-indigenous language query terms overall.

International (non-Aotearoa) clients exhibited different information retrieval behaviours from national (Aotearoa) clients. Further, the behavioural differences noted between international and national reo Māori clients were not the same as the behavioural differences noted between international and national English language clients. These observations suggest that the international indigenous language client uses unique information retrieval behaviours, behaviours that are unlike the national indigenous language retrieval behaviours and the international non-indigenous language behaviours.

The default language setting of the digital library affected indigenous and non-indigenous language clients differently. When the default language was set to the indigenous language, usage requests in that language increased and when the default language was set to the non-indigenous language, usage requests in the non-indigenous language increased. These results were not surprising but what was unexpected was the results that occurred when the interface language did not match the client's preferred interface language. When the default interface language was set to the indigenous language there was a reduction in the number of usage requests by non-indigenous language clients and the non-indigenous language clients reduced by a third. When the default language of the interface was set to the non-indigenous language there was a similar reduction in the number of usage requests in the indigenous language but no reduction in the number of indigenous language clients. These characteristics suggest that the indigenous language client was a bilingual client who could tolerate the default language of the interface being in either language. However, the non-indigenous language client, perhaps not being bilingual, was often intolerant of the default language not matching their preferred interface language and, therefore, a third chose not to use the digital library in this situation.

8.6 Bilingual Users

The interface was often used bilingually to enhance the information retrieval activity.

There appeared to be a deliberate strategy by some clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries to use the interface bilingually to enhance information retrieval. These clients submitted the greatest number of queries per session (though by definition these sessions were long), altered the Some/All query parameter the most frequently and submitted the greatest mix of indigenous and non-indigenous language query terms. Queries submitted by these clients appeared to be the most effective in locating the required documents.

The information retrieval behaviours displayed by bilingual clients of the Māori Niupepa digital library were very similar to those by bilingual clients of the Hawaiian Nūpepa digital library. What was not similar was the number of bilingual usage requests made to each of the digital libraries; 3,921 bilingual usage requests were made to the Māori Niupepa digital library which represents 3.2% of the total Niupepa usage requests and 50,359 bilingual usage requests to the Hawaiian Nūpepa digital library representing 25.4% of the total Nūpepa usage requests.

The design of the interface was perhaps the primary reason for significantly more bilingual usage requests to the Hawaiian Nūpepa digital library than to the Māori Niupepa digital library. The Hawaiian Nūpepa interface allowed clients to toggle the interface language between Hawaiian and English on each interface page while the Māori Niupepa interface only offered this toggle facility on the home page. Māori Niupepa clients wanting to change the interface language had to return to the home page or access the preferences page. This extra action appears to have been too much trouble for many of the potential bilingual clients because the Māori Niupepa interface was rarely used bilingually.

8.7 Interface Usability and Design

The TLA and user feedback showed some usability issues with the interface design of both digital libraries. Most of the critiquing appeared to be from non-indigenous clients; indigenous clients appeared less concerned with usability issues and/or more reluctant to mention problems which they had encountered. There are three potential reasons why indigenous clients appeared unwilling to mention interface problems: because indigenous clients often had more interest in the texts which led them to disregard the flaws in the interface; because indigenous clients were less experienced in online interfaces and, consequently, did not have high expectations; or perhaps the unwillingness is part of the indigenous culture where belittling a person or their work is not seen as constructive. Whatever the reasons were, less negative feedback was noted from indigenous language clients.

Although the usability evaluation of these digital libraries was not intended as part of this research, it is apparent that if certain usability issues were addressed then the use of the indigenous language is likely to increase. Some of the usability issues are detailed in this section to assist other interface designers intending to design bilingual interfaces and/or interfaces that use indigenous languages.

Translations

Translations should be unambiguous, accurate and relevant to the context. The te reo Māori translation for the link “About English Abstracts” was listed as “He Rāpopototanga Reo Pākehā”. This phrase could be misconstrued to mean “a summary in English”. TLA indicated the reo Māori version of this link was requested at an unusually high rate, which suggests confusion and, thus, ambiguity. Some of the Chinese users also noted some confusion in the Chinese translations of newspaper chapter numbers and, subsequently, chose not to use the

interface in Chinese. These examples show that inaccurate translations can influence language usage in digital libraries.

The site should be translated completely. Some multilingual users mentioned that if they noted a site partially translated they would switch the interface language to the dominant language.

Bilingual sites do not need direct translations. TLA indicated that clients who used the site in the indigenous language were most interested in indigenous language documents. Clients who used the site in the non-indigenous language were most interested in the non-indigenous language documents. Consequently, an indigenous language version of the interface should focus on highlighting and directing clients to the indigenous language content in the first instance. In a similar manner, perhaps a non-indigenous language version of the interface should highlight and direct clients to the non-indigenous language content in the first instance.

Content Description and Format

The different types of content should be clearly described. Observation of clients using the Māori Niupepa digital library indicated that they were often unsure if they were reading newspaper documents or English abstracts about the newspaper documents. They were also unaware that the English abstracts only summarised te reo Māori documents. Users were observed searching the English abstracts for English language texts, only to be disappointed when the results returned only links to te reo Māori documents.

The abstracted text documents need disclaimers. Many clients were unaware that the extracted text documents had been generated by an OCR process that was not 100% accurate. While errors in the extracted texts of documents were sometimes noted and commented on by clients, many clients were unaware that the full size image of the document was an exact replica and, thus, contained no extraction errors.

Image files stored in pdf format with a high resolution can easily be enlarged to assist with reading. The pdf format also allowed the document pages to be easily printed.

A list of titles predominantly in one language should also include translations for the benefit of clients who do not understand that language. Even with the interface set to English, the series list (see *Figures 3-4* and *6-4*) was perceived as foreign by clients who did not understand the indigenous language. Parallel translations of the titles could assist those clients.

Query Activity

The number of documents that are searchable should be made clear. Many clients assumed that all digital library documents were searched when a query was submitted. However, 98.6% were searched in the Māori Niupepa digital library and only 6.4% in the Hawaiian Nūpepa digital library.

The more searchable documents the better. Both TLA and the user interviews indicated that an important feature of the digital libraries was the ability to undertake full text searching. Clients of the Hawaiian Nūpepa digital library expressed an eagerness for more document pages to become available for full text searching as quickly as possible.

Query parameters and query operators should function correctly. The query parameters and query operators appeared to act incorrectly or inconsistently in one of the digital libraries. Consequently, clients could not rely on these to work correctly and use of the operators was subsequently abandoned.

Search results should contain preview images of the returned documents. A preview image of the result, especially when text surrounding the query term was included, allowed clients to decide quickly if the document was relevant to the query. Results could be scanned much more quickly and the search activity was more efficient when preview documents were available. Clients also suggested

that when a document was displayed following a query, the document should be automatically scrolled to the query term position and the query term highlighted.

The user interviews revealed how surprisingly poor digital library users were at spelling. A spell checking function incorporated into the query term input box would improve the submission of queries. If the spell checker worked with both the indigenous and non-indigenous languages it would be even more useful.

Navigation and Layout

Navigation should be clear and logical. Users commented that at times they were unsure where they were in the digital library and how they could return to a previous screen. Buttons and icons should afford their purpose. User feedback suggested that clients were not aware of the purpose of some buttons and icons simply by looking at them and that many were unaware that some buttons were, in fact, buttons; they were written as text with no sign that they were clickable hyperlinks.

The screen layout should be consistent with common practice. Users mentioned that navigation would be easier if the buttons were located in a similar position to that used on other web sites; the home button at the top left and the help button at the top right. The button locations should be consistent throughout the site.

A language toggle button on every page of the interface would assist clients to interact with the digital library bilingually.

8.8 Results Requiring Further Investigation

The Transaction Log Analysis used in Chapters 3-6 yielded a large amount of information. While all relates to indigenous language usage in bilingual digital libraries and, as such, is relevant to this research, much of this information is

beyond the scope of this thesis. Further work is required in many research areas to complete the analysis of indigenous language usage in digital libraries.

For example, an early section in Chapters 3-6 discussed results from how clients explored but then did not go on to use a digital library. Data was provided about the lengths of these exploratory sessions, the apparent read times of interface pages and which interface pages clients requested before deciding not to go on and use the digital library. This data is relevant to investigations into why certain online resources/web pages are used while others are just briefly explored. The differences shown in indigenous language and non-indigenous language exploratory sessions would give insights into online exploratory differences between indigenous and non-indigenous people. Further investigation of these results is required by researchers in the field of human-computer interaction (HCI).

Another topic requiring further investigation is the propensity for indigenous language clients to request documents published in certain years (see *Figures 3-23* and *6-16*). Why were these particular years favoured? Because of historical events that happen in those years, because those were the years that most of the periodicals were published, because those were the documents with the most extracted text formats or because certain authors were publishing certain content in those years? Further investigation is required by historians, sociologists and even librarians to find answers to these questions.

The data generated about interface language switching (see §3.4.7, §4.3.5, §5.3.5 and §6.4.6) also requires further investigation. It appears that indigenous language clients alter the language of the interface a lot less often than non-indigenous language clients. This could suggest that many of the indigenous language clients are bilingual or multilingual. The point at which the interface language is switched within the sessions could suggest a frustration by clients with that particular interface. A consistent level of switching by a user group through a session could indicate a deliberate strategy of using the interface bilingually. Applied linguistics research could provide answers about why these switching behaviours occurred.

The research described in this thesis examines activity in the year 2005. The same methods could be used to examine activity recorded in 2004, 2006, and in subsequent years. A longitudinal study of this nature would provide valuable information about how indigenous language use was evolving in online information systems.

8.9 Claims Revisited

In the introduction the following claim was made:

This thesis charts indigenous language usage by clients of a digital library and demonstrates how digital libraries can provide an environment to support the use of indigenous languages.

Subsequent chapters used Transaction Log Analysis (TLA) to chart indigenous language usage in the Māori Niupepa and Hawaiian Nūpepa digital libraries. The TLA has shown how the indigenous language was used by clients in these digital libraries and some settings that appear to encourage indigenous language usage. Feedback from digital library clients has proposed why the indigenous language was used and has suggested some design principles that would further promote indigenous language usage.

This thesis has shown that indigenous languages were used in digital library environments in 2005. The indigenous language use was discretionary; that is, it took place without coercion. Indigenous languages were successfully used to retrieve information in a modern information system. Using te reo Māori and the Hawaiian languages as examples, this research has shown how digital libraries can provide an environment to support the use of indigenous languages.

8.10 Final Comments

This thesis began with reo Māori poetry from a whakawātea that emphasised the importance of the te reo Māori. It is appropriate to also end with reo Māori poetry, this time taken from a hakaⁱⁱ entitled, *Tōku Reo! Tōku Reo!* ('My Language! My Language!')ⁱⁱⁱ The words are an encouragement and a challenge to Māori people to take responsibility for te reo Māori survival into the future. The challenge has been laid, it is up to Māori to seize this challenge and assist with the revitalisation of te reo Māori. No translation is given.

E tātou e te ao Māori, taringa whakarongo!
Ko te reo te moko o te hinengaro!
Ki te mate tō tātou reo
Ehara kē i te mea me mate
He kore kē ia nō tātou
I aro atu, i hautoa kia ora tonu ai
Nei rā te mānuka kua takoto
Kawe ake! Kawe ake!
Hei!

ⁱⁱ haka: a Māori dance with postures.

ⁱⁱⁱ This haka was composed by Tīmoti Kāretu and first performed in 1994 by the Te Whare Wānanga o Waikato kapahaka (performing arts) group at national competitions in Te Hāwera.

References

- Alosius, J. (2007). Students Enrolled in Public Provider by Ethnic Group in 2005 (Spreadsheet Provided by Ministry of Education). Wellington: Data Management Unit, Ministry of Education.
- Apperley, M. D., Keegan, T. T., Cunningham, S. J., & Witten, I. H. (2002). Delivering The Māori Newspapers on the Internet. In J. Curnow, N. Hopa & J. McRae (Eds.), *Rere Atu Taku Manu! Discovering History Language and Politics In The Māori Language Newspapers*. (pp. 211–236). Auckland: Auckland University Press.
- Barbour, R. H., Cunningham, S. J., & Ford, G. (1993). Maori Wordprocessor for Indigenous New Zealand Young Children. *British Journal of Educational Technology*, 24(2), 114–124.
- Bedford, R., Didham, R., Ho, E., & Hugo, G. (2004). Maori Internal and International Migration at the Turn of the Century: An Australasian Perspective. *New Zealand Population Review*, 30(1/2), 131–142.
- Benton, R. A. (1978). *Can the Maori Language Survive?* Wellington: New Zealand Council for Educational Research.
- Benton, R. A. (1981). *The Flight of The Amokura*. Wellington: New Zealand Council for Educational Research.
- Benton, R. A. (1996). Making the Medium the Message: Using an Electronic Bulletin Board System for Promoting and Revitalising Maori. In M. Warschauer (Ed.), *Telecollaboration in Foreign Language Learning* (pp. 187–204). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i.
- Benton, R. A. (1997). *The Māori Language: Dying or Reviving?* Wellington: New Zealand Council for Educational Research.

- Blecic, D., Bangalore, N., Dorsch, J., Hendersen, L., Koenig, M., & Wellar, A. (1998). Using Transaction Log Analysis to Improve OPAC Retrieval Results. *College and Research Libraries*, 59(1), 39–50.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design-For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: Jossey-Bass.
- Buszard-Welcher, L. (2001). Can the Web Help Save My Language. In L. Hinton & K. Hale (Eds.), *The Green Book of Language Revitalisation in Practise* (pp. 331–345). San Diego, California: Academic Press.
- Cartwright, J., Chantiny, M., Hori, J., & Peacock, K. (2000). The Digital Landscape: Hawiian Newspapers, War Records, and Trust Territory Image Repository of the University of Hawaii [Electronic Version]. *First Monday*, 5, 15. Retrieved 26 June 2006 from:
www.firstmonday.org/issues/issue5_6/cartwright/index.html.
- Chapin, H. G. (1996). *Shaping history the role of newspapers in Hawai'i*. Honolulu: University of Hawai'i Press.
- Chau, M., Fang, X., & Sheng, O. (2005). Analysis of the Query Logs of a Web Site Search Engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363–1376.
- Crystal, D. (2000). *Language Death*. Cambridge: Cambridge University Press.
- Cunliffe, D. (2004). Promoting Minority Language Use on Bilingual Websites. *Mercator Media Forum*, 7(1), 42–53.
- Dallimore, G. (1990). He Arahi, He Tohu o Nga Pepa Maori: A bibliography of Maori Newspapers, 1840–1900. Unpublished manuscript.
- Durie, M. (2004). Understanding Health and Illness: Research at the Interface Between Science and Indigenous Knowledge. *International Journal of Epidemiology*, 33(5), 1138–1143.
- Dyson, L. E., Hendicks, M., & Grant, S. (2006). *Information Technology and Indigenous People*. Hershey PA: Information Science Publishing.
- Fishman, J. A. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon: Multilingual Matters.
- Gordon, R. G. (2005). *Ethnologue: Languages of the World*, Fifteenth edition. Electronic Version. Retrieved January 2007 from: www.ethnologue.com.
- Grenoble, L. A., & Whaley, L. J. (2006). *Saving Languages. An Introduction to Language Revitalization*. New York: Cambridge University Press.

- Hall, E. T. (1983). *The Dance of Life: The Other Dimension of Time*. New York: Anchor Press/Doubleday.
- Harlow, R. (1987). *A Word-List of South Island Maori*. University of Auckland: Linguistic Society of New Zealand.
- Jansen, B. (2006). Search log analysis: What it is, what's been done, how to do it. *Library and Information Science Research*, 28(3), 407–432.
- Jansen, B., & Spink, A. (2005). How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs. *Information Processing and Management*, 42, 248–263.
- Jansen, B., Spink, A., & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries On the Web. *Information Processing and Management*, 36(2), 207–227.
- Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. J. (2000). A Transaction Log Analysis of a Digital Library. *International Journal on Digital Libraries*, 3(2), 152–169.
- Keegan, P. (1996, 14-17 January). *Initiatives in Maori Networking*. Proceedings of the 6th Biennial Conference of the New Zealand Computers in Education Society, Hamilton, New Zealand (pp. 35–52).
- Keegan, T. T., Apperley, M., Cunningham, S. J., & Witten, I. (2001, 23-26 July). *The Niupepa Collection: Opening the Blinds on a Window to the Past*. Proceedings of the ICHIM01 International Cultural Heritage Informatics Meeting Milan, Italy (pp.347–356).
- Keegan, T. T., & Cunningham, S. J. (2005a, 7–11 June). *Language Preference in a Bi-language Digital Library*. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, Colorado, USA. (pp. 174–175).
- Keegan, T. T., & Cunningham, S. J. (2005b, May 26–28, 2005). *What Happens If We Switch The Default Language of a Website*. Proceedings of the First International Conference on Web Information Systems and Technologies, Miami, Florida (pp. 263–269).
- Kāretu, T. (1996). *Ngā Waiata me ngā Haka a te Kapahaka o Te Whare Wānanga o Waikato (He Whakahoutanga)*. Hamilton: University of Waikato.
- Laws, M. R. (1995). *A Computer Assisted Māori Language Laboratory. The Design, Development and Implementation of an integrated package of Māori language course material in a computerised environment—is it*

- possible and practical*. Proceedings of the Inaugural NAMMSAT Conference, Auckland University.
- Laws, M. R. (2001). *Māori Language Integration in the Age of Information Technology: A Computational Approach*. Unpublished PhD Thesis, University of Otago, Dunedin.
- Metge, J. (1967). *The Maoris of New Zealand*. London: Routledge & K. Paul.
- Nichols, D. M., Witten, I. H., Keegan, T. T., Bainbridge, D., & Dewsnip, M. (2005). Digital Libraries and Minority Languages. *New Review of Hypermedia and Multimedia*, 11(2), 139-155.
- Nielsen NetRatings. (2005). Reaching NZ Māori Online. Retrieved January 2007 from: www.nielsen-netratings.com/pr/pr_050420_newzealand.pdf.
- Office of the Domain Name Commissioner. (2005). Second Level Domains Electronic Version. Retrieved April 2006 from: www.dnc.org.nz.
- Peters, T. A. (1993). The History and Development of Transaction Log Analysis. *Library Hi Tech*, 11(2), 41-66.
- Pool, I. (1991). *Te Iwi Māori: A New Zealand Population Past, Present & Projected*. Auckland, New Zealand: Auckland University Press.
- Reedy, T. M. (2000). Te Reo Māori: The Past 20 Years and Looking Forward. *Oceanic Linguistics*, 39(1), 157-169.
- Russell Bernard, H. (1992). Preserving Language: Computers Can Be a Tool For Making the Survival of Languages Possible. *Cultural Survival Quarterly*, 16(3), 15-18.
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(1), 6-12.
- Smith, L. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. Dunedin: University of Otago Press.
- Stauffer, R. (2006). Personal communication, June 13, 2006, Honolulu, Hawai'i.
- Statistics New Zealand. (2006). 2006 Census Data. Retrieved January 2007 from: www.stats.govt.nz/census/2006-census-data/.
- Tan, P., & Kumar, V. (2002). Discovery of Web Robot Sessions Based on Their Navigational Patterns. *Data Mining and Knowledge Discovery*, 6(1), 9-35.

- Te Puni Kōkiri. (1998). *The National Māori Language Survey: Te Mahi Rangahau Reo Māori*. Wellington: Te Puni Kōkiri (Ministry of Māori Development).
- Te Puni Kōkiri. (2001a). Māori Access to Information Technology. Retrieved January 2007 from: www.tpk.govt.nz/publications/.
- Te Puni Kōkiri. (2001b). Te Māori i Ngā Rohe. Retrieved April 2006 from: www.tpk.govt.nz/publications/.
- Te Puni Kōkiri. (2004). Te Reo Māori i te Hapori—Māori Language in the Community [Electronic Version]. Retrieved January 2007 from: www.tpk.govt.nz/publications/.
- Walker, R. (2005). *Quality Assurance in Tertiary Education from a Māori (Indigenous) Perspective*. Proceedings of the International Network for Quality Assurance Agencies in Higher Education (INQAAHE), Wellington, New Zealand.
- Wang, P., Berry, M., & Yang, Y. (2003). Mining Longitudinal Web Queries: Trends and Patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743–758.
- Warschauer, M. (2000). Language Identity and the Internet. In B. Kolko, L. Nakamura & G. Rodman (Eds.), *Race in Cyberspace* (pp. 151-170). New York: Routledge.
- Warschauer, M., & Donaghy, K. (1997). Leokī: A Powerful Voice of Hawaiian Language Revitalization. *Computer Assisted Language Learning*, 10(4), 349–361.
- Williams, H. W. (1985). *A Dictionary of the Maori Language* (Seventh ed.). Wellington: Government Printing Office.
- Wilson, W. H. (2006). Personal Communication. In T. T. Keegan (Ed.). Hilo: Ka Haka 'Ula O Ke'elikolani College of Hawaiian Language, University of Hawai'i at Hilo.
- Witten, I. H., & Bainbridge, D. (2002). *How to Build a Digital Library*. San Francisco, CA: Morgan Kaufmann.

Appendix A Pearl Scripts Summary

Perl scripts were written to process the raw web logs. A description of their use in generating Transaction Log Analysis data is given in §2.5. The Perl scripts filtered the web logs to remove inappropriate requests, classified the requests into sessions and generated statistical information. The output of the Perl scripts were text files which were entered into spreadsheets so that graphs and further statistical information could be generated. The Perl scripts were written by Gareth Judsen under supervision in 2005/2006 while Gareth was a student of the Computer Science Department at the University of Waikato. Summaries of the individual Perl scripts are tabled below.

Many of the Perl scripts require a specific directory structure to run correctly. This structure consists of:

- <base directory>—directory containing the initial raw log and may have any name
- <base directory>/Results—directory to receive the output text files
- <base directory>/Results/Reading Times—directory to receive the reading files
- <base directory>/Session Data—directory to receive the sessions files

A series of tables below describe each of the Perl scripts, and lists the input files required and the output files produced by the Perl scripts. This is followed by a list of data and statistics produced by the Perl scripts.

To obtain a copy of the Perl scripts contact the author, Te Taka Keegan at the Computer Science Department, University of Waikato, Aotearoa.

Name:	process-raw-logs.pl
Description:	Used for processing raw log to remove unwanted CGI arguments and unwanted requests, such as requests to incorrect collection, non-English and reo Māori requests, requests from web robots, etc. Outputs summary information indicating how many requests were removed and why, lists web robots found by agent string in robot_agents.txt and web robots found by IP address in robot_ips.txt.
Size of Script:	319 lines of code
Input Files:	Raw log (can have any name) e.g. <base directory>/RawLog.txt
Output Files:	<base directory>/PROCESSED-RawLog.txt (i.e. the input file with the string 'PROCESSED-' added to it)
Additional Requirements:	The script takes the input file and year for the requests as a command line argument, in that order. The script requires 'logs.pl' 'robot_agents.txt' and 'robot_ips.txt'; all should be located in the same directory as this script.
<i>Table A-1: Summary of script process_raw_logs.pl.</i>	

Name:	create-sessions.pl
Description:	Creates session files from the processed logs. Produces a results file summarising the counts of the various sessions.
Size of Script:	387 lines of code
Input Files:	<base directory>/processedRawLog.txt
Output Files:	<base directory>/Session Data/Zargument-singles.txt <base directory>/Session Data/Zargument-exploratory.txt <base directory>/Session Data/Zargument-usage.txt <base directory>/Session Data/IP-singles.txt <base directory>/Session Data/IP-exploratory.txt <base directory>/Session Data/IP-usage.txt <base directory>/Results/General Session Data.txt
Additional Requirements:	The script takes the input file as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.
<i>Table A-2: Summary of create-sessions.pl.</i>	

Name:	split-sess-lang.pl
Description:	Divides the session files into languages of session files. Generates switching data and files.
Size of Script:	509 lines of code
Input Files:	<base directory>/Session Data/Zargument-exploratory.txt <base directory>/Session Data/Zargument-usage.txt <base directory>/Session Data/IP-exploratory.txt <base directory>/Session Data/IP-usage.txt
Output Files:	<base directory>/Session Data/Zargument-exploratory-Maori.txt <base directory>/Session Data/Zargument-exploratory-English.txt <base directory>/Session Data/Zargument-usage-Maori.txt <base directory>/Session Data/Zargument-usage-English.txt <base directory>/Session Data/Zargument-usage-Bilingual.txt <base directory>/Session Data/IP-exploratory-Maori.txt <base directory>/Session Data/IP-exploratory-English.txt <base directory>/Session Data/IP-usage-Maori.txt <base directory>/Session Data/IP-usage-English.txt <base directory>/Session Data/IP-usage-Bilingual.txt <base directory>/Results/Zargument-exploratory-switching.txt <base directory>/Results/Zargument-usage-switching.txt <base directory>/Results/IP-exploratory-switching.txt <base directory>/Results/IP-usage-switching.txt
Additional Requirements:	The script takes the <base directory> as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.
Table A-3: Summary of split-sess-lang.pl.	

Name:	analyse-single-types.pl
Description:	Produces data about single request sessions.
Size of Script:	282 lines of code
Input Files:	<base directory>/Session Data/Zargument-singles.txt
Output Files:	<base directory>/Results/Single Request Data.txt
Additional Requirements:	The script takes the <base directory> as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.
Table A-4: Summary of analyse-single-types.pl.	

Name:	session-stats.pl
Description:	Produces large amounts of statistics about sessions. Produces files that contain apparent read times.
Size of Script:	1107 lines of code
Input Files:	<base directory>/Session Data/Zargument-exploratory.txt <base directory>/Session Data/Zargument-usage.txt <base directory>/Session Data/Zargument-exploratory-Maori.txt <base directory>/Session Data/Zargument-exploratory-English.txt <base directory>/Session Data/Zargument-usage-Maori.txt <base directory>/Session Data/Zargument-usage-English.txt <base directory>/Session Data/Zargument-usage-Bilingual.txt
Output Files:	<base directory>/Results/Exploratory Session Data.txt <base directory>/Results/Usage Session Data.txt <base directory>/Results/Exploratory Sessions in Maori Data.txt <base directory>/Results/Exploratory Sessions in English Data.txt <base directory>/Results/Usage Sessions in Maori Data.txt <base directory>/Results/Usage Sessions in English Data.txt <base directory>/Results/Bilingual Usage Sessions Data.txt <base directory>/Results/Reading Times/PageRead-ExploratoryMi.txt <base directory>/Results/Reading Times/PageRead-ExploratoryEn.txt <base directory>/Results/Reading Times/DocRead-UsageMi.txt <base directory>/Results/Reading Times/PageRead-UsageMi.txt <base directory>/Results/Reading Times/DocRead-UsageEn.txt <base directory>/Results/Reading Times/PageRead-UsageEn.txt <base directory>/Results/Reading Times/DocRead-UsageBi.txt <base directory>/Results/Reading Times/PageRead-UsageBi.txt
Additional Requirements:	The script takes the <base directory> as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.

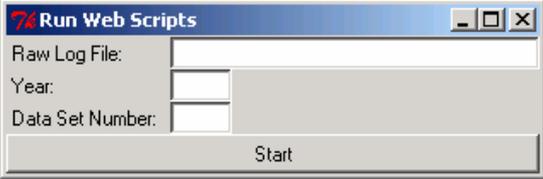
Table A-5: Summary of session-stats.pl.

Name:	logs.pl
Description:	A utility script containing functions for reading, manipulating and outputting raw cgi logs like those produced by the greenstone web interface. Equivalent functions are also available for the logs processed using the functions in this script.
Size of Script:	416 lines of code
Input Files:	
Output Files:	
Notes:	Required by all Perl scripts except 'renameFile.pl', 'renameLengths.pl' and 'gui.pl'.

Table A-6: Summary of logs.pl.

Name:	analysis.pl
Description:	A utility script for use in session analysis. Functions are available for reading in sessions from session logs, creating summary statistics, getting time and query term data.
Size of Script:	379 lines of code
Input Files:	
Output Files:	
Additional Requirements:	Requires logs.pl
Notes:	Required by all Perl scripts except 'logs.pl', 'process-raw-logs.pl', 'analyse-single-types.pl', 'renameFile.pl', 'renameLengths.pl' and 'gui.pl'.

Table A-7: Summary of analysis.pl.

Name:	gui.pl
Description:	<p>An interface to make using the set of scripts easier.</p> <p>Allows users to drag and drop the raw log, enter the year of the log, enter the data set number for this log then start to run all scripts.</p> 
Size of Script:	96 lines of code
Input Files:	
Output Files:	
Additional Requirements:	User must specify year of the script and what dataset it will have. All the other logs mentioned here must be in the same directory as this script.
<i>Table A-8: Summary of gui.pl.</i>	

Name:	renameFiles.pl
Description:	Simple script run to rename all the results files to a format specified by the file names.txt.
Size of Script:	76 lines of code
Input Files:	names.txt (in same directory as the log) (The files that are altered are in <base directory>/Results/) <base directory>/Results/General Data.txt <base directory>/Results/General Session Data.txt <base directory>/Results/Single Request Data.txt <base directory>/Results/Exploratory Session Data.txt <base directory>/Results/Usage Session Data.txt <base directory>/Results/Exploratory Sessions in Maori Data.txt <base directory>/Results/Exploratory Sessions in English Data.txt <base directory>/Results/Usage Sessions in Maori Data.txt <base directory>/Results/Usage Sessions in English Data.txt <base directory>/Results/Bilingual Usage Sessions Data.txt <base directory>/Results/Zargument-usage-switching.txt
Output Files:	(Assuming we pass a base number of 1) <base directory>/Results/ 1.1 – General Data.txt <base directory>/Results/ 1.2 – General Session Data.txt <base directory>/Results/ 1.3 – Single Request Data.txt <base directory>/Results/ 1.4 – Exploratory Session Data.txt <base directory>/Results/ 1.5 – Usage Session Data.txt <base directory>/Results/ 1.6 – Exploratory Session in Maori Data.txt <base directory>/Results/ 1.7 – Exploratory Session in English Data.txt <base directory>/Results/ 1.9 – Usage Sessions in Maori Data.txt <base directory>/Results/ 1.10 – Usage Sessions in English Data.txt <base directory>/Results/ 1.11 – Bilingual Usage Sessions Data.txt <base directory>/Results/ 1.12 – Switching Data.txt
Additional Requirements:	The base number must be passed to the script as a command line argument.
Notes:	files are renamed as follows: <base number>.<secondary number> - <description>.txt Where <base number> is a number passed to this script and <secondary number> and <description> are specified in names.txt
Table A-9: Summary of renameFiles.pl.	

Name:	renameLengths.pl
Description:	Simple script which is run to rename all the length files to a format specified in the script.
Size of Script:	56 lines of code
Input Files:	(All the files to be renamed are in /Results/Reading Times/) <base directory>/Results/Reading Times/PageRead-ExploratoryMi.txt <base directory>/Results/Reading Times/PageRead-ExploratoryEn.txt <base directory>/Results/Reading Times/DocRead-UsageMi.txt <base directory>/Results/Reading Times/PageRead-UsageMi.txt <base directory>/Results/Reading Times/DocRead-UsageEn.txt <base directory>/Results/Reading Times/PageRead-UsageEn.txt <base directory>/Results/Reading Times/DocRead-UsageBi.txt <base directory>/Results/Reading Times/PageRead-UsageBi.txt
Output Files:	(Assuming we have a base number of 1) <base directory>/Results/Reading Times/1.6 - PageRead - ExploratoryMi.txt <base directory>/Results/Reading Times/1.7 - PageRead - ExploratoryEn.txt <base directory>/Results/Reading Times/1.9 - DocRead - UsageMi.txt <base directory>/Results/Reading Times/1.9 - PageRead - UsageMi.txt <base directory>/Results/Reading Times/1.10 - DocRead - UsageEn.txt <base directory>/Results/Reading Times/1.10 - PageRead - UsageEn.txt <base directory>/Results/Reading Times/1.11 - DocRead - UsageBi.txt <base directory>/Results/Reading Times/1.11 - PageRead - UsageBi.txt
Additional Requirements:	The base number must be passed to the script as a command line argument.
Notes:	files are renamed as follows: <base number>.<secondary number> - <doc OR page> - <Session><Lang>.txt Where <base number> is a number passed to this script, <secondary number> is specified in this script and should relate to the session stats file numbers—that is the usage English stats file should have the same numbers as the doc and page read files relating to it. <doc OR page> is either DocRead or PageRead depending on whether this is a page read length or document read length file. <Session> is either Usage or Exploratory depending on what the type of session file is. <Lang> is either En, Mi, or Bi depending on whether this session log file was an English (En), reo Māori (Mi) or Bilingual (Bi) session.
Table A-10: Summary of renameLengths.pl.	

The Perl scripts *analyse-single-types.pl* and *session-stats.pl* generated the majority of the data used in the Transaction Log Analysis. Examples of the data and statistics generated include:

i) General Data

- Counts of requests in reo Māori and in English.
- Counts of unique users in reo Māori and in English.
- Counts of total requests and total unique users.
- Counts of pages requested; home page, about page, explanation page, help page, preferences page, gsdl page and unknown pages.
- Counts of documents requested from search, series, date and unknown.
- Counts of document types requested, text, full image, preview and other.

ii) Session Length Summary Data

- Statistical data on length of session in time includes; mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on length by requests includes; mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.

iii) Session Length Listing

- Listing of session counts by length of sessions in minutes.
- Listing of session counts by length of sessions in requests.

iv) Chronological Data

- Counts of requests per month.
- Counts of requests per day of week.
- Counts of requests per hour of the day.
- Counts of sessions per month.
- Counts of sessions per day of week.
- Counts of sessions per hour of the day.

v) Requests per Client

- Listing of Z arguments by count of sessions and counts of requests.
- Listing of IP addresses by count of cookies.
- Listing of IP addresses by count of requests.

vi) Session Query Data

- Counts of queries and query pages.

- Statistical data on queries per session includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Counts of queries that used search operators, e.g. Boolean, quotation etc.
- Counts of queries that changed the query arguments, including maximum query requests, search type (Some/All) and case matching.
- Counts of sessions that had no queries, 1 query, 2 queries, 3 queries, 4 queries, 5 queries and 6 or more queries.
- Statistical data on number of terms per query includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Count of queries with 1 query term, 2 query terms and 3 or more query terms
- Statistical data on number of documents viewed after a query includes; mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on number of text documents, full size image documents and preview images viewed after a query.
- Statistical data on time (in seconds) between document views after a query includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on time (in seconds) between text documents viewed after a query, between full size images viewed after a query and between preview images viewed after a query.

vii) Listing of Query Terms

- Listing of query terms by number of times submitted and number of unique users submitting.

viii) Documents Viewed

- Listing of documents viewed by count of requests and count of unique user.

ix) Times Between Requests and Views

- Statistical data on time between successive page requests includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on time between successive document requests includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.

Appendix B Hawaiian Query Terms

The query terms submitted to the Hawaiian Nūpepa search engine were collated but not listed, in Chapter 6 Hawaiian Language Usage of the Niupepa Digital Library. The query terms were not listed because it was felt that the analysis of these terms was beyond the scope of the thesis. However, the query terms are listed in this appendix for the benefit of those who would like to analyse them.

In total there were 58,846 query terms submitted to the Hawaiian Nūpepa digital library in 2005. There were 18,478 query terms used in 11,232 queries undertaken in usgHaw sessions, 22,026 query terms used in 14,966 queries undertaken in usgEn sessions, and 14,342 query terms used in 9,428 queries undertaken in usgBi sessions.

Two methods can be used to determine the most popular query terms. The first method is simply a count of each time a term, or string of terms, was submitted to the search engine. The second method is a count of unique users who submitted the search term or search string. Terms are displayed sorted by unique user as this reduces the effect of users who are making repeat searches on the one term and gives a better indication of the variety of topics for which clients were searching.

The most common query terms submitted are displayed in *Table B-1*. The number of times that the term was requested is in square brackets, the number of unique users requesting that term is in curved brackets.

Most Common Hawaiian Nūpepa Query Terms in 2005

	usgHaw	usgEn	usgBi
1	hula [30] (17)	Pahau [58] (24)	liliuokalani [23] (11)
2	bush [23] (16)	kalawaia [39] (20)	hula [31] (9)
3	aloha [24] (15)	kamehameha [24] (18)	kamehameha [13] (7)
4	liliuokalani [24] (13)	kahoalii [55] (17)	kealoha [7] (7)
5	kanikau [33] (12)	bush [34] (15)	Liliuokalani [11] (5)
6	welina [30] (10)	naholowaa [25] (15)	annexation [8] (5)
7	kohola [28] (10)	Kamehameha [25] (14)	kalakaua [7] (5)
8	mokaulele [18] (10)	hula [26] (12)	silva [6] (5)
9	Kaupo [36] (9)	kaaawa [14] (12)	kalawaia [47] (4)
10	kaiulani [20] (9)	manoa [13] (11)	Wond [34] (4)
11	Pele [17] (9)	liliuokalani [12] (11)	Makuaole [30] (4)
12	kumulipo [14] (9)	kaneohe [12] (10)	kaumualii [18] (4)
13	maui [13] (9)	holua [24] (9)	welina [16] (4)
14	kaleponi [35] (8)	nahinu [19] (9)	napulou [15] (4)
15	kealohi [25] (8)	napulou [23] (8)	Queen Emma [14] (4)
16	kaao [18] (8)	Kolomoku [23] (8)	Kamehameha [14] (4)
17	palapala sila nui [16] (8)	hana [14] (8)	mary kealoha [9] (4)
18	kalo [14] (8)	mahele [9] (8)	kamapuaa [9] (4)
19	kaluaikoolau [14] (8)	kawaihau [22] (7)	kauhane [9] (4)
20	Kamapuaa [14] (8)	Purdy [19] (7)	heenalu [8] (4)
21	mookuauhau [13] (8)	kamakau [18] (7)	molokai [7] (4)
22	mano [9] (8)	naki [15] (7)	mookuauhau [7] (4)
23	Keauhou [25] (7)	heenalu [13] (7)	hana [6] (4)
24	Kahekili [23] (7)	kekuhaupio [12] (7)	kaaawa [6] (4)
25	Kawelo [19] (7)	hawaiian language [12] (7)	bush [5] (4)
26	kawelo [12] (7)	kahele [12] (7)	
27	Liliuokalani [11] (7)	kanikau [10] (7)	
28	koolau [9] (7)	overthrow [9] (7)	

Table B-1: Query terms with [number of requests] and (number of unique users) submitted by usgHaw, usgEn and usgBi clients to the Hawaiian Nūpepa digital library.

Appendix C Interview Forms

The interview procedure is described in §7.3. The interviews began with an explanation of the purpose of the research. The participants were then shown, and had explained to them, a Bill of Rights and a Research Consent Form which they were asked to sign if they were happy to be interviewed. This consent form was also signed by the interviewer/researcher.

The Bill of Rights form, shown in *Figure C-1*, consists of two parts. The first part, written only in te reo Māori, greets the participants in a traditional manner and expresses gratitude for taking time to be involved with this research. The second part, written only in English, lists nine rights that each participant was entitled to throughout all stages of the interview procedure.

The Research Consent Form, shown in *Figure C-2*, gives some background information about the research and the interview procedure. It then details what the participant has agreed to and provides a place for the signatures of the participant and the researcher. Two copies of the form were signed; one was retained by the participant, the other by the researcher.

Kaupapa Rangahau: te Kohikohinga Niupepa/Nūpepa

He hōnore, he korōria ki runga.
He maunga rongō ki te mata o te whenua.
He whakaaro pai tētehi ki tētehi.

Rire, rire, haul Pai mārire.

E ngā mana e ngā reo e ngā tapu e noho atu nā, tēnei te tuohu te tuku mihi tēnā koutou katoa.

E tuku mihi ana hoki ki a koe kua whai wā koe ki te noho mai ki te āwhina i te kaupapa, ki te whakamahi i te Kohikohinga Niupepa, ki te kōrero hoki mō ētehi āhuatanga kua kitea e koe. Hei aha? Hei arahi hei āwhina i ngā kōkiri whakamua o tō tātou reo rangatira i roto i ngā kaupapa hāngarau pēnei nā. Nō reira ka nui rawa te mihi, tēnā hoki koe.

Kāti. Ka kite atu e rua ngā kaupapa kei runga i ngā whārangi nei. Tuatahi he whakamārama i ōu mōtika, arā ngā *Participant's Rights*, tuarua he puka whakaaetanga māu mō te mahi rangahau nei, arā te *Research Consent Form*.

Tēnā koa, pānuitia ngā kōrero nei, ā, mehemea e pai ana hainatia te puka whakaaetanga.

Mehemea, he pātai, he āwangawanga, kōrero mai koa.

Nā tō pononga noa nei,

nā Te Taka Keegan
Pūkenga, Kairangahau, Tāri Rorohiko
Te Whare Wānanga o Waikato

Research Participant's Bill of Rights

The following is a list of your rights if you participate in this usage evaluation of Niupepa/Nūpepa.

As a research participant, you have the right:

1. To be treated with dignity and respect at all times.
2. To choose not to be involved in this evaluation.
3. To withdraw from this evaluation at any time.
4. To know that all information you provide in this evaluation is anonymous, and will be kept strictly confidential.
5. To be fully and clearly informed of all aspects of the research prior to becoming involved in it.
6. To enter into clear, informed, and written agreement with the researcher prior to becoming involved in this activity. You should sense NO pressure, explicit or otherwise, to sign this contract.
7. To be treated with honesty, integrity, openness, and straightforwardness in all phases of the website usage evaluation.
8. To expect that your personal welfare is protected and promoted in all phases of the research, including knowing that no harm will come to you.
9. To access any information that has been collected about your evaluation and to remove this information if you are no longer comfortable with it being used in this research.

Figure C-1: Bill of Rights for Research Participants.

Research Consent Form

** Please take the time to read and understand this form.*

Research Project Title: Niupepa/Nūpepa Website Usage Evaluation

Researcher: Te Taka Keegan

Evaluation Purpose: The purpose of this evaluation is to investigate bilingual usage of the Niupepa and Nūpepa Websites. Web site log analysis has highlighted different usage characteristics of people who access the site in Māori/Hawaiian compared with people who access the site in English. The intention of this evaluation is to try to determine why some of those differences are occurring.

Participant Recruitment and Selection: Potential users of the Niupepa/Nūpepa website are solicited by the researcher and colleagues and invited to participate in this evaluation.

Procedure: This evaluation consists of three distinct stages. First you will be introduced to the purpose and procedure of the Niupepa/Nūpepa website evaluation. Next you will be asked to use the Niupepa/Nūpepa website to investigate some topic of interest to you. Finally you will be asked some questions about your experience with using the Niupepa/Nūpepa website.

Explaining the evaluation and allowing you to read the information is expected to take about 15 minutes, using the website is expected to take about 20 minutes, and the final questioning stage is expected to take about 20 minutes. Consequently the evaluation should take approximately 1 hour.

It is your choice as to whether this evaluation is conducted in te reo Māori or in English.

Data Collection: Your usage of the website and the questions you are asked afterwards will be videoed and recorded to DVD. This is to provide an accurate record of your activity and responses. This DVD will not be publicly available. It will not be privately available outside the purposes of this research without prior written consent from the participants. The DVD will be kept secure in my office, in G.2.07 at the University of Waikato. On completion of the research the DVDs will be destroyed.

Confidentiality: Confidentiality and participant anonymity will be strictly maintained. You will not be identified in any publication or dissemination of research findings.

Likelihood of Discomfort: There is no likelihood of discomfort or risk associated with participation.

Finding out about Results: You can find out about the results of the study by contacting the researcher, Te Taka Keegan by email (tetaka@waikato.ac.nz) after December 2006.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in this research and agree to participate as a participant. In no way does this waive you legal rights nor release the researchers from their legal and professional responsibilities. You are free to withdraw from the study at any time. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please ask the researcher.

Participant

Date

Investigator/Witness

Date

**You may keep a copy of this consent form for your records for future reference should you need it.*

Figure C-2: Research Consent Form.

1 Introduction

‘Tōku reo, tōku ohooho
Tōku reo, tōku māpihi maurea
Tōku reo, tōku whakakai marihi
Aue! Aue! Aue! Te mamae ki taku taonga e!

My language, my cherished possession
My language, my object of affection
My language, my precious adornment
Aue! Aue! Aue! What pain I feel at the loss
of something so valuable. (Kāretu 1996)ⁱ

These words describe the importance of te reo Māoriⁱⁱ to the Māori people. They are words from a whakawātea,ⁱⁱⁱ titled, *He Aha Rā Kei Te Tau o Taku Ate?* (What Is This That is Tearing Me Apart?).^{iv} The whakawātea raises the serious question of how te reo Māori can survive, and challenges Māori to ‘get your act together’ so that the language can become revitalised again.

The research described in this thesis suggests that the revitalisation of te reo Māori and other indigenous languages can be assisted by their use in digital libraries.

ⁱ See Kāretu (1996, pp. 75–76).

ⁱⁱ The Māori language.

ⁱⁱⁱ An item performed in kapahaka (performing arts group) competitions where the performers exit the stage. Whakawātea literally means to leave clear.

^{iv} This whakawātea was composed by Tīmoti Kāretu and first performed in 1988 by the Te Whare Wānanga o Waikato kapahaka at national competitions in Whāngārei.

Why Investigate Digital Libraries?

Communication is increasingly occurring and information is increasingly being provided, electronically. From the advent of the radio and telephone, through the widespread use of television and video, to the propagation of computers and the Internet, the media by which societies communicate and transmit information have taken many electronic forms. Languages are likely to survive and prosper if they can be used in these new electronic environments.

One particular electronic environment with the potential to provide new situations where languages can be used, and thus assist in language revitalisation, is a digital library. Witten and Bainbridge (2002, pp xxvi) define a digital library as a focused collections of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organisation and maintenance. A digital library's potential to assist use of a language is twofold. First it has the potential to make available a large amount of information about a language and/or in a language readily available to speakers of that language. With suitable content in a particular language it has the potential to attract a large number of those language speakers. Second, it has the potential to offer an interface in any written language, thus offering speakers of that language an opportunity to use their language in a modern electronic environment.

If a digital library with an indigenous language interface makes available indigenous language material, and that material is highly sought after, then there is the potential for the digital library to be *used* in the indigenous language. Revitalisation of indigenous languages is likely to occur if indigenous language speakers choose to use their indigenous languages in new environments such as digital libraries.

Why Use the Term Indigenous Language?

Indigenous language is a phrase used in this thesis to signify the traditional language that is spoken by indigenous people, for example te reo Māori in

Aotearoa/New Zealand.^v Indigenous languages are languages that originate in a particular country or locale and are usually spoken by people who have an affinity to that locale. While many languages could be classified as indigenous languages, this research uses the term indigenous language to describe traditional languages that are not in a dominant position and are in threat of extinction. The precarious survival of these languages is the main reason that this research is so important.

The decision to use the term indigenous is not taken lightly as there are an abundance of terms that have been used to describe similar language research in the computing and internet environments. These terms include native, aboriginal, heritage, local, community, non-state, endangered, threatened, disadvantaged, lesser-used, less commonly taught and minority. While the language primarily studied in this research, te reo Māori, could be described using all of these adjectives, the term indigenous is preferred as it does not imply a language from a specific timeframe and it does not depict a language in a marginalised or negative manner. The term indigenous does portray a language with mana,^{vi} characteristically mana whenua^{vii} and is the term that is preferred in recent publications by Māori scholars (Durie, 2004; Smith, 1999; Walker, 2005).

Finally, the term indigenous language is used as a reminder to the reader that indigenous language users are, in the most part, indigenous users. Their purposes, their methods and their characteristics in using their indigenous language are indigenous purposes, indigenous methods and indigenous characteristics. This analysis of indigenous language usage by clients^{viii} of digital libraries will take into account indigenous perspectives and indigenous viewpoints.

^v Aotearoa/New Zealand is a modern phrase used to describe the country commonly known as New Zealand. For the sake of brevity Aotearoa/New Zealand will be reduced to Aotearoa.

^{vi} Mana: prestige, authority, held in great esteem.

^{vii} Mana Whenua: in this case mana that is derived from having a long existence in or long association with a certain locale.

^{viii} Client is the term used in this research to describe users who log on to a digital library and use it through the web interface.

1.1 Thesis Statement

This thesis charts indigenous language usage by clients of a digital library and demonstrates how digital libraries can provide an environment to support the use of indigenous languages.

This claim is that digital libraries provide a modern electronic medium where the use by choice, that is use without coercion or encouragement of an indigenous language can occur. The claim is substantiated in two parts. The first explains *how* indigenous languages are used by clients of digital libraries and what settings encourage indigenous language use. The second part suggests *why* indigenous languages are used by clients of digital libraries and gives characteristics of digital library environments that promote the usage of the indigenous language.

1.2 Background

The indigenous language that receives the most scrutiny in this research is te reo Māori of Aotearoa/New Zealand. A brief description is provided of the origin of the word Māori, the decline and rise of the Māori population and te reo Māori use is outlined. This is followed by a summary of why information technology, including digital libraries, is so important to indigenous languages. The digital library comprehensively analysed in this research is the Niupepa^{ix} digital library. This section concludes by introducing the Niupepa digital library and explaining why this particular digital library is appropriate for an investigation of indigenous language use.

^{ix} Literally ‘the Newspapers’, a collection of newspapers published in Māori or for a Māori audience between 1840 and 1933. For a full description see §1.2.3 and §3.1.1.

1.2.1 The Decline of Te Reo Māori

Te reo Māori is threatened with extinction (Benton, 1978; Te Puni Kōkiri, 1998). The indigenous language of Aotearoa is a minority language amongst its indigenous people. This sub-section summarises the decline of te reo Māori emphasising the importance of this research in seeking ways to assist in its revitalisation.

The word Māori is commonly used to describe the indigenous people of Aotearoa.^x Traditionally the term māori means ‘usual, ordinary or normal’ (Williams, 1985). The phrase tāngata māori^{xi} was initially used by Māori to distinguish themselves from the early European explorers to Aotearoa. Later, European settlers adopted the shorter version of the term, Māori, to describe the indigenous people of Aotearoa and over time this term became globally accepted (Metge, 1967).

Te reo Māori has been classified as being in the Eastern-Polynesian group of languages (Harlow, 1987). It resides in the Tahitic sub-class of Central Eastern Polynesian languages, along with Austral (French Polynesia), Penryn (Cook Islands), Rakahanga-Manihiki (Cook Islands), Rarotongan (Cook Islands), Tahitian (French Polynesia), Tuamotuan (French Polynesia) and Rapa (Easter Islands) (Gordon, 2005).

Estimates suggest the Māori population numbered over 100,000 prior to European contact and then declined to as low as 42,000 in the late 1800’s, primarily due to introduced diseases and land alienation (Pool, 1991). As the Māori population decreased the European population increased. In 1840 the Māori population outnumbered the European population by 40:1 but by the end of the 19th century the European population outnumbered the Māori population by 16:1. From the 1900s the Māori population decline ceased and from the 1920s a gradual regeneration began because of an improved immunity to diseases and an

^x The indigenous people of the Cook Islands are also referred to as Māori, though more commonly as Cook Island Māori in Aotearoa.

^{xi} Tāngata is people, māori is normal or common; hence tāngata māori means the normal people.

improvement in social economic status (Pool, 1991). Results from the 2006 census indicate that there were 565,329 Māori living in Aotearoa (Statistics New Zealand, 2006). The 2006 census also showed that 14.6% of Aotearoa's resident population were Māori, 67.6% were European, and the remaining 17.8% consisted of other ethnicities.

The use of te reo Māori in Aotearoa declined with the decline in Māori population in the 19th century. A lower proportion of Māori people meant there was less need for the European settlers to learn te reo Māori. The passing and rigorous enforcing of the Native Schools Act in 1867, which emphasised the use of English and excluded the use of te reo Māori in the classroom and school grounds, was a contributing factor for the language shift that was to occur in Māori communities. The ban on te reo Māori in schools, together with the rapid urbanisation of Māori in the 1940s and 1950s and the pepper potting policy (Benton, 1997), led to a massive language shift to English amongst Māori people (Te Puni Kōkiri, 2004). While pockets of rural communities still existed where te reo Māori was the prominent language, research undertaken in the late 1970s suggested that reo Māori was destined to vanish in the near future (Benton, 1978, 1981).

The imminent threat to te reo Māori gave rise to many initiatives seeking to redress its impending extinction. Such initiatives included the establishment of Te Kōhanga Reo (The Language Nest—te reo Māori pre-schools) in 1982; Kura Kaupapa Māori (Reo Māori primary schools based on Māori ideology) in 1985; Te Taura Whiri i te Reo Māori (Māori Language Commission) in 1987; Te Māngai Pāho (Māori Broadcasting Agency) in 1993, and a host of other initiatives^{xii} and efforts by both community and government agencies.

While the revitalisation efforts gave rise to a lot of optimism, a large national reo Māori survey undertaken in 1995 indicated that te reo Māori was still in a critical state. Only 21.7% of Māori had a fluency above a low level (Te Puni Kōkiri, 1998). Recent figures from the 2006 census list 131,613 Māori (24.4% of Māori) who indicated that they could speak te reo Māori (Statistics New Zealand, 2006).

^{xii} For a general description of Māori revitalisation efforts see Reedy (2000).

However, these figures do not indicate the complexity of reo Māori that is being spoken, the ability to read and write in Māori, or if intergenerational transmission is occurring. Intergenerational transmission is recognised as being crucial for language survival (Fishman, 1991). The outlook for the survival of te reo Māori as a language of everyday use is by no means assured.

To be revitalised as a language of everyday use, te reo Māori must be used by choice, without coercion or pressure, by its speakers in everyday situations. Opportunities in all endeavours of life must be created where the discretionary use of te reo Māori will occur. These include but are not limited to domestic settings, social events, sporting activities, education and business discourse, and all other opportunities both verbal and non-verbal when language is used for communication. One particular endeavour, a component of which is investigated in this thesis, is the use of language in information technology.

1.2.2 Information Technology & Indigenous Languages

The importance of information technology to indigenous languages has been described by Hale Kuamo‘o’s^{xiii} current director Keiki Kawai‘ae‘a, who states, “...In order for Hawaiian to feel like a real living language, like English, it needs to be seen, heard and utilized everywhere, and that includes the use of computers” (cited in Warschauer & Donaghy, 1997, pp. 352-353).

There are many ways that information technology, in the form of computing and networking devices, can be used to assist languages threatened with extinction. Information technology can be used to document and archive language information; create a wide range of multi-media resources for language learning (and preservation); connect communities of language speakers and thus create new environments where the threatened language can be used; as a source of new identity for speakers and to instil further pride in the language. David Crystal, in

^{xiii} A detailed description of Hale Kuamo‘o is given in §6.1.2.

his book *Language Death* discusses language revitalisation factors and postulates "An endangered language will progress if its speakers make use of electronic technology" (Crystal, 2000, p. 141).

Using information technology to assist with the survival of languages is not a new idea. In 1987 the Native Literacy Centre was established in Oaxaca, Mexico, to enable (Mexican) Indian people to read, write and publish in their own languages using computers (Russell Bernard, 1992). In 1994 Hawaiian educators established Leokī, a graphic bulletin board system installed on computers throughout immersion schools in Hawai‘i. Entirely in Hawaiian, the system facilitated email, chat, conferences, announcements and provided online dictionaries and newspapers (Warschauer & Donaghy, 1997).

The potential of information technology was also realised early by technologists working with te reo Māori. A bilingual computer system designed to support the teaching and learning of te reo Māori was created by the Kōhanga Reo movement in 1986 (Laws, 2001). A computer bulletin board called Te Wahapū was set up by the New Zealand Council for Educational Research in 1990. With its operating language in te reo Māori, Te Wahapū was designed to facilitate discussion on issues related to te reo Māori and Māori education, and to provide access to a database of new technical terms in Māori (Benton, 1996; P. Keegan, 1996). A word processor called Tā Kupu was designed in 1992 at the University of Waikato (Barbour, Cunningham, & Ford, 1993) in order to teach written reo Māori skills to Māori children. In 1993 staff at the University of Otago began to integrate reo Māori course material into a computerised laboratory so that reo Māori students could use computers to assist with their learning of te reo Māori (Laws, 1995).

Over the past two decades information technology has been used to such an extent that it has prompted the editors of a recent book of collected articles, entitled *Information Technology and Indigenous Peoples*, to state "...the use of information technology within indigenous communities and by indigenous peoples is no longer an issue of debate but a proven fact. It is no longer a dream of the future but a reality of today" (Dyson, Hendicks, & Grant, 2006, p.314).

The Internet, part of the spectrum of information technology, has the potential to make the most impact on indigenous languages. However, Grenoble & Whaley (2006) have suggested that local communities in many parts of the world are still a long way from having internet access. A report based on a survey undertaken in 2000 stated that only 35% of Māori respondents had ever used the Internet (Te Puni Kōkiri, 2001a). It appears that four years later this figure had more than doubled; a report released by Nielsen NetRatings stated that 72% of all Māori had access to the Internet (Nielsen NetRatings, 2005).

There has been considerable speculation that the Internet, and one particular component of the Internet, the World Wide Web (Web), can assist in language revitalisation. In an article titled *Can the Web Help Save My Language* Buszard-Welcher (2001) examined endangered language web sites of Native North Americans. She discovered 50 endangered language web sites with varying amounts of language material available. While stating that few of these sites had substantial amounts of language content, she did suggest that we are only beginning to realise the potential of the Web for language maintenance and revitalisation. Warschauer (2000) studied a number of Hawaiian language initiatives using the Internet and suggested that interacting in the Hawaiian language gave the students (of Hawaiian) an opportunity to explore and strengthen their sense of individual and collective Hawaiian identity. He suggested the most important role of the Internet is not its impact on transmission but its impact on will. He quoted Nancy Hornberger (Warschauer, 2000, p. 13), “language revitalisation is not about bringing a language back, it’s about bringing it forward!” Warschauer goes on to say that people will fight to maintain their language when it is not only an important part of their grandparents’ past but also of their own future. Cunliffe (2004) suggested that the presence of minority languages in the medium of the Web will become as important to language survival as having a presence in traditional media.

By having a web interface to a digital library, the digital library becomes available to all people who are able to connect to the Web. Consequently the potential and promise of the Web to assist in language revitalisation is also applicable to indigenous language digital libraries with a web interface. One such indigenous

language digital library is the Niupepa digital library, which is introduced in the following sub-section.

1.2.3 The Niupepa Digital Library

In 1996 the Alexander Turnbull Library^{xiv} made available a collection on microfiche called 'Māori Newspapers', or 'Niupepa 1842–1933'. The collection had been sourced by scouring libraries of Aotearoa for historic newspapers that were either published in te reo Māori or for a Māori audience. The collection comprised 407 microfich, covering 35 titles and 17,700 facsimiles of newspaper pages. Approximately 54% of the newspapers were written only in te reo Māori, 44% were written in parallel reo Māori and English texts and the remaining 2% were written solely in English.

The Niupepa collection contains a large amount of historical information and opinions embodying Aotearoa's early encounter history. The collection is also a rich source of an earlier reo Māori, providing a much needed text source for scholars, teachers and students of te reo Māori.

Having this collection available as microfiche was a breakthrough for researchers as it meant they could now access all the newspapers in the Niupepa collection from the one source. However, searching the collection was tedious and time consuming. The Computer Science Department at the University of Waikato resolved to redress this situation by digitising the texts, undertaking OCR (Optical Character Recognition) to allow full text searching and making the collection available online using digital library software^{xv} with a web interface.

In 2001 the Niupepa collection was made available online using the New Zealand Digital Library's (NZDL) Greenstone software^{xvi} at www.nzdl.org/niupepa.

^{xiv} An entity within the National Library of New Zealand specialising in heritage documents

^{xv} For a comprehensive explanation regarding the process of delivering the Niupepa on the Web, see T.T. Keegan et al (2001) and Apperley et al (2002).

^{xvi} For an explanation of Greenstone software see Witten & Bainbridge (2002).

The Niupepa digital library is designed with multilingual functionality with a switchable interface where the default language of the web page is normally set to reo Māori. A button on the home page allows the default language of the interface to be switched between English and reo Māori. The digital library provides three facilities for accessing the newspapers: full text search, browse by individual title and browse by date. For a full description of the Niupepa web site interface, including screenshots, see §3.1.1.

There are five reasons why the Niupepa digital library is an ideal choice to undertake usage analysis of an indigenous language. First, because the contents of the Niupepa Collection are important and valuable to Māori people it is likely that Māori people would seek to access the digital library without coercion. As the Niupepa digital library is readily available on the Web, at no charge, the likelihood of use is increased. Second, as a large proportion of the Niupepa Collection's content is in te reo Māori it is also likely to be accessed by users who are literate in te reo Māori. Third, as the interface was often set to te reo Māori it is likely that discretionary use of te reo Māori could occur. Fourth, having the ability to easily switch the interface language between the indigenous and non-indigenous language provides access to the collection by users who have no or little proficiency in te reo Māori.

The fifth reason was ease of access. The digital library hosting the Niupepa Collection was administered and maintained by the Computer Science Department at the University of Waikato, the same department where this research was undertaken. Access to the web logs and the digital library configuration files was therefore easily obtained.

1.3 Thesis Approach

Use of an indigenous language was charted by a study undertaken on the Niupepa digital library. The study was conducted from the start of January 2005 to the end of December in that year. It involved collecting all requests made to the Niupepa web site and analysing them to determine how clients used the digital library in

the indigenous language (te reo Māori). The study revealed three types of clients with significantly different usage characteristics. There were clients who preferred to use the Niupepa digital library with the interface language set to te reo Māori, clients who preferred to use the digital library with the interface language set to English and clients who preferred to use both te reo Māori and English interface language settings in their information retrieval activities.

Analysis was concentrated on the activities of te reo Māori clients. Results showed that use of te reo Māori increased if the default language of the web site was set to te reo Māori. Further analysis was undertaken to determine if the usage of the indigenous language was restricted to the locality of that indigenous population. This analysis indicated that usage of the Niupepa digital library in te reo Māori was not restricted to clients from Aotearoa; a significant number of te reo Māori international clients were using the Internet to connect and use the Niupepa digital library.

To ensure that the results were not peculiar to te reo Māori, a similar study was conducted on another indigenous language digital library. For the year of 2005 all requests to the Hawaiian Nūpepa^{xvii} digital library^{xviii} were collected and analysed as described above. These results also showed different usage characteristics amongst clients who used the interface in different languages. The results indicated that the indigenous language was used more in the Hawaiian Nūpepa digital library than it was in the Māori Niupepa digital library.

Usability studies were undertaken to determine why indigenous languages were used by clients of digital libraries and what characteristics of this environment encouraged indigenous language usage. The usability studies were undertaken in two forms—a web questionnaire and client interviews. The web questionnaire was placed on the Māori Niupepa web site for 250 days in 2006. It received 60 responses, mostly from clients who preferred to use the indigenous language (te

^{xvii} The terms Niupepa and Nūpepa are both Polynesian words for newspaper. Niupepa is the term commonly used for the Māori newspaper collection whereas Nūpepa is the term commonly used for the Hawaiian newspaper collection.

^{xviii} A full description of the Hawaiian Nūpepa digital library is given in §6.1.

reo Māori). Sixteen interviews were conducted with clients who used the digital libraries in te reo Māori, in Hawaiian and in English. The interviews involved a usage observation section and a question/answer section. The user feedback was analysed to reveal that clients used the indigenous language because of the content of the digital library, the default language setting of the digital library and because of a personal desire to use the indigenous language as often as possible.

1.4 Thesis Outline

This thesis is arranged into eight chapters. The first chapter claims that digital libraries can be used to assist in the usage, and subsequent revitalisation, of indigenous languages. The relevance of this research has been highlighted by describing the current position of te reo Māori, the importance of information technology to the revitalisation of te reo Māori and the ways in which the Niupepa digital library is suited to indigenous language usage analysis.

Transactional Log Analysis (TLA) is the method used in this research to give quantitative accounts of *how* indigenous languages are used to access information by clients of digital libraries. Chapter 2 provides a comprehensive explanation of how TLA is used to uncover reo Māori usage in the Niupepa collection. The methodology formulated in Chapter 2 is then applied in the next four chapters to determine reo Māori usage by clients of the Niupepa collection (Chapter 3), the difference between national and international reo Māori usage (Chapter 4), how the default language setting affects reo Māori usage (Chapter 5), and Hawaiian language usage by clients of the Hawaiian Nūpepa digital library (Chapter 6).

Chapter 7 uses feedback from clients to determine *why* clients prefer to use indigenous languages in bilingual digital libraries. Feedback was obtained from use of a web questionnaire and client interviews. Although the sample sizes were too small to make general claims, the questionnaire and interviews suggest that clients use the indigenous language because it matches the content, it is the default

language of the digital library and they have a desire to use the indigenous language.

Chapter 8 summarises the key findings, raises some interface design issues and asserts that digital libraries can provide an environment to support the use of indigenous languages.

2 Transaction Log Analysis

This thesis claims that digital libraries can provide an environment to support the use of indigenous languages. This claim is substantiated by defining *how* indigenous languages are used by clients of digital libraries. Transaction Log Analysis (TLA) is the method used to generate quantitative data on indigenous language usage in a digital library environment.

This chapter begins by introducing TLA and explaining some methodological decisions made in this research. The TLA undertaken on the Niupepa digital library is then described. The process involves collecting raw datasets, preparing the datasets, defining user sessions, and then analysing the sessions to determine if and how indigenous language is being used in the digital library. This TLA process is described in a simple but comprehensive manner so that the methods used can be repeated and re-applied to other bilingual or multilingual information systems.

2.1 Introduction

A transaction log is a detailed record (log) of all interactions (transactions) between a system and users (clients) of that system. This record is usually stored as a file on the system and analysis of this file is termed Transaction Log Analysis (TLA). TLA has been described as "the study of electronic recorded interactions between on-line information retrieval systems and the persons who search for information found in those

systems" (Peters (1993) cited in Jansen (2006, p. 408)). TLA can be performed by commercially available web log analysis tools; such tools were considered too inflexible and the results too limited for this particular research. Data mining techniques (also known as web usage mining or web log mining) can also be used to identify usage patterns from web logs. Data mining techniques were also considered too limited for the wide-ranging results required for this analysis of indigenous language users interacting with a digital library.

Many publications have used TLA to analyse user behaviours.ⁱ For example, Blečić et al (1998) employed TLA to reveal that many experienced difficulty with basic searching techniques when using an OPAC (Online Public Access Catalogue) system. The OPAC system's introductory screens were simplified and standardised, and key word searching was given more prominence. A second TLA determined that users' searching behaviours had improved significantly as a result of the system changes. Jansen et al (2000) used TLA to analyse queries from the Excite search engine. They reported that web users were uncomfortable with advanced methods of searching, did not use Boolean operators and did not frequently browse the results beyond the first page or so.

TLA was used by Jones et al (2000) to investigate user activity on the digital library of Computer Science Technical Reports Collection. A general description of computer science researchers' behaviours in a digital library revealed that sessions were usually very short, with few queries and the queries themselves were very simple. Chau et al (2005) used TLA to investigate whether search behaviour on a general purpose search engine, such as Excite or AltaVista, was different from search behaviour on a search engine restricted to one particular web site. They found that, in general, web users behaved similarly in each case, although there were differences in the number of queries per session, the query terms used and the query topics searched.

ⁱ A comprehensive review of transaction log analysis studies is beyond the scope of this thesis. However, Peters (1993) provides an early review and Jansen & Spink (2005) a more recent review.

TLA was used by Silverstein et al (1999) to analyse the query log of the AltaVista Search engine. The authors reported that actual web users differed significantly from the users assumed in standard information retrieval literature. Queries were short, seldom modified and the users mostly looked at the first 10 results only. They did note, however, that their methodology did not exclude data generated by non-human activity, i.e. by web robots. A technique for determining activity generated by web robots was proposed by Tan & Kumar (2002). They suggested that patterns of activity could be used to create robot classifiers that would accurately identify robot activity in transaction logs. For the TLA undertaken in this thesis, web robot activity was detected primarily by examining the User Agent and IP Addresses of the clients (see §2.3.2.2), with transaction activity also examined for unusual behaviour that could be attributed to web robots.

Jansen (2006) suggests importing transaction log data into a relational database to undertake analysis, a method popular with other researchers (Chau, Fang, & Sheng, 2005; Wang, Berry, & Yang, 2003). However, text processing scripts were used in this research to classify the data into sessions and then the sessions into language types; it was decided to continue to use text processing scripts to generate the descriptive statistics.

The TLA undertaken on the Niupepa digital library in this thesis is significant from two perspectives. First, a large amount of usage information was generated. Over 40 distinct categories of data were examined, representing significantly more categories than previous TLA studies. Second, this TLA focuses on analysing how indigenous languages are used by clients of digital libraries. Apart from preliminary analysis undertaken on the Niupepa collection by the author (Keegan & Cunningham, 2005a, 2005b; Keegan et al, 2006; Nichols et al, 2005), the TLA methodology has not otherwise been used for indigenous language usage analysis.

2.2 Collecting Datasets

The first step when undertaking Transaction Log Analysis is to collect the transaction log data. This data was collected from three sources because there were three web servers making the Māori Niupepa digital library and the Hawaiian Nūpepa digital library available to the Web in 2005.

A web server named Puka, located in Hamilton, Aotearoa,ⁱⁱ served all Aotearoa requests to the `www.nzdl.org` domain. Another web server, named Thor, located in Lethbridge, Canada, contained a mirror of the NZDL web site and served all international requests to the `www.nzdl.org` domain. The third web server, Ulukau, located in Hilo, Hawai‘i, served all requests to the `www.nupepa.org` domain. All three web servers were running the open source Apache HTTP Server software and all three were using the Greenstone Digital Library software to log similar web request information.

Each discrete request submitted by a client’s browser to one of the web servers was recorded in the web server’s log file as a single transaction. These transactions were previously termed *hits* but now are more commonly known as *requests*. The transaction log data was collected by these servers into large text files containing a defined set of details about every request received by the web servers.

The requests captured by the three servers mentioned above were copied into six web log datasets depending on certain criteria. *Table 2-2* gives a summary of the six datasets. The criteria which define these datasets are described in the following sub-sections.

ⁱⁱ Also known as Aotearoa/New Zealand and New Zealand.

2.2.1 Dataset 1 NZDL 2005

Dataset 1 consisted of all web browser requests to the `www.nzdl.org` domain name during 2005. It was gathered by combining all the requests recorded in the Puka web log with all the requests recorded in the Thor web log. In total there were 1,133,543 requests which, as a text file, required 1.16 gigabytes of disk space. As Dataset 1 contained all requests to all digital libraries housed at the NZDL web site the Niupepa digital library data was extracted by filtering out all the non-Niupepa requests. This activity is described in §2.3.2. A comprehensive analysis of Dataset 1 was undertaken to determine how reo Māori was used by clients of the Niupepa digital library. This analysis is detailed in Chapter 3.

2.2.2 Dataset 2 National and Dataset 3 International

Datasets 2 and 3 were collected so that a comparative analysis could be undertaken on the Niupepa digital library. Reo Māori usage from national (Aotearoa) users was compared with reo Māori usage from international (non-Aotearoa) users. This analysis was undertaken to determine if digital libraries can be used both nationally and internationally to assist in the revitalisation of indigenous languages.

As previously mentioned, the digital libraries available through the NZDL web site were served by two web servers. This occurred from 2003 to 2006 and involved Puka, located in the Computer Science Department at the University of Waikato, Aotearoa, and Thor, located at the Department of Mathematics and Computer Science at the University of Lethbridge, Alberta, Canada.

Requests to `www.nzdl.org` were originally received by the web server Puka. Lines of code in the Apache configuration file located on Puka pre-processed all requests to the NZDL web site. If the IP Address of the requestor did not have

.nz as a first level domain name then the request was redirected to be handled by the web server Thor. Consequently, all the requests from national users, as indicated by the user's IP Address, were served by Puka, the web server located in Aotearoa. All other requests were served by Thor, the web server located in Canada.

Creating a national (requests from Aotearoa) dataset and an international (requests from outside of Aotearoa) dataset was straightforward. The Puka web server was serving requests known to be originating from within Aotearoa so the 2005 web log from this server became Dataset 2 National Requests. The Thor web server was serving all other requests so the 2005 web log from this server became Dataset 3 International Requests.

Dataset 2 National 2005 consisted of 291,140 requests in a 268 megabyte text file. Dataset 3 International 2005 consisted of 842,403 requests stored in a 890 megabyte text file. A comparative analysis of Dataset 2 and Dataset 3 was undertaken to identify national and international differences in how te reo Māori was used by clients of the Niupepa digital library. This comparative analysis is described in Chapter 4.

2.2.3 Dataset 4 Default Reo Māori and Dataset 5 Default English

Previous TLA studies have indicated that clients do not like to alter the default settings on a digital library interface (Jones et al, 2000). The 2006 national census (Statistics New Zealand, 2006) also indicated that almost all (98.5%) of the Māori population that could speak a language, could speak English. Consequently, one could hypothesise that if the interface to the Niupepa digital library was set to English the reo Māori literate bilingual clients would use the interface in English. Alternatively, if the interface was set to reo Māori they would use the interface in te reo Māori. Dataset 4 and Dataset 5 were collected to test this theory and, thus,

determine how setting the default language of the interface influenced the usage of te reo Māori by clients of the Niupepa digital library.

The default language setting of the Niupepa interface is set by altering the CGI (Common Gateway Interface) language argument on the Niupepa default home page link. When the language argument on the default home page link is set to reo Māori, `l=mi`, as in the following example;

```
www.nzdl.org/cgi-bin/library?a=p&p=about&c=niupepa&l=mi&nw=utf8
```

then all the texts, commands and buttons on the user interface of the Niupepa web site are displayed in te reo Māori. When the language argument on the default home page link is set to English, `l=en`, as in the following example;

```
www.nzdl.org/cgi-bin/library?a=p&p=about&c=niupepa&l=en&nw=utf8
```

then all the texts, commands and buttons on the user interface of the Niupepa web site are displayed in English. The default home page link is stored in a macro file called *nzdlhome.dm* which is located on the web servers. This link is used when web browsing software requests the default Niupepa home page.

If a client subsequently alters the interface language setting of the Niupepa web site, the altered language will stay in effect until the client reloads the home page. The home page URL does not include any CGI arguments; thus, a reload on this page will cause the predefined default settings, as in the examples listed above, to be reinstated. If a client bookmarks a page of the NZDL and the bookmark includes the CGI `L` Argument, then a request for the bookmarked page will not invoke the default language setting of the Niupepa interface.

At the start of 2005 the default language of the Niupepa interface was set to reo Māori. Throughout 2005 this default language was altered on a monthly basis. This frequency was chosen to cause as little disruption to the user as possible, yet seasonal usage activity would still be captured in both datasets. In the months of January, March, May, July, September and November the default language argument of the Niupepa interface was set to reo Māori. In the months of February, June, August, October and December the default language argument of

the interface was set to English. The dates and times that the default language was altered are listed in *Table 2-1*.

Set on PUKA		Set on THOR		Default Lang.
1 st January	00:00 am	1 st January	00:00 am	reo Māori
4 th February	1:32 pm	3 rd February	5:35 pm	English
1 st March	9:50 am	28 th February	1:50 pm	reo Māori
1 st April	9:15 am	31 st March	2:42 pm	English
<i>Unknown</i>	<i>Unknown</i>	<i>Unknown</i>	<i>Unknown</i>	reo Māori
2 nd May	8:50 am	1 st May	2:48 pm	reo Māori
9 th June	9:49 am	8 th June	3:51 pm	English
1 st July	3:56 pm	30 th June	9:59 pm	reo Māori
1 st August	9:17 am	31 st July	3:19 pm	English
1 st September	8:54 am	31 st August	2:56 pm	reo Māori
3 rd October	10:00 am	2 nd October	3:03 pm	English
1 st November	10:45 am	31 st October	2:48 pm	reo Māori
2 nd December	9:59 am	1 st December	2:02pm	English

Table 2-1: The dates, times and (new) default languages that were set on the Niupepa web site in 2005

Some time during April the default language of the Niupepa interface was (by mistake) set to reo Māori. This setting may have occurred as a result of some software upgrades that happened at that time. Because it is not known for certain exactly when in April this default language setting was changed all April requests were excluded from Dataset 4 and Dataset 5.

The default language was set to te reo Māori for 198 days (54.2%) and to English for 137 days (37.5%). There were 30 days (8.2%) when, because of uncertainty in the default language setting, the web log data was not used.

Dataset 4 Default Reo Māori contains all requests received by the Niupepa web site when the default interface language was set to te reo Māori (as outlined in *Table 2-1*). Dataset 4 Default Māori consists of 598,260 requests and is stored in

608 megabyte file. Dataset 5 Default English contains all requests received by the Niupepa web site when the default interface language was set to English (as outlined in *Table 2-1*). Dataset 5 Default English consists of 407,145 requests and is stored in a 417 megabyte file. Both datasets contain requests from national (Aotearoa) users and international (non-Aotearoa) users.

A comparative analysis of Datasets 4 and 5 was undertaken to identify whether altering the default language of the Niupepa interface influenced the usage of te reo Māori by clients of the Niupepa digital library. This comparative analysis is detailed in Chapter 5.

2.2.4 Dataset 6 Hawaiian Nūpepa 2005

Dataset 6 Nūpepa consisted of all requests to the `www.nupepa.org` domain name in 2005. This domain is responsible for serving the Hawaiian Nūpepa digital library. This digital library is similar to the Māori Niupepa digital library in that it made available a large collection of historic newspapers written in an indigenous language using Greenstone software. However, there were some interface differences between the Hawaiian Nūpepa digital library and the Māori Niupepa digital library. These are listed in *Table 6-1*. Because of these differences, minor changes needed to be made to the scripts that generated the Transaction Log Analysis statistics for Dataset 6. For example, requests in Hawaiian were collated as opposed to requests in te reo Māori; counts were undertaken of middle sized documents as opposed to just large and preview documents; data about preference pages and help pages could not be collected as these pages were not available with the Hawaiian Nūpepa interface.

On the 3rd of December 2005 the Nūpepa web site was moved onto a new web server. The new web server had a default setting that utilised its ‘web performance cache’. This default setting had the effect of making all requests appear in the log as if they came from the IP Address 128.171.24.14; the web server’s IP address.

This situation continued until mid January when the web site administrators noticed what was happening and reset the default setting so that the originating requestors' IP addresses were correctly recorded in the web logs. While normal activity continued it had all been recorded incorrectly as being from the one IP address. The method used to define sessions does not rely on IP addresses; thus, it can be ascertained that over this period; there were 3,174 single requests, 133 exploratory session requests, and 17,384 usage sessions requests, a total of 20,691 requests.

However, there is a small chance that the incorrect recording of the IP addresses compromised the session data. If two new users made a request to the Nūpepa collection within the same second they both would have been allocated the same unique identifier, meaning both users would appear in the logs as the same entity. Such a situation would compromise the data. Although the chances of this situation occurring were very small it was decided to remove all requests from the IP address 128.171.24.14. Removing these requests had the effect of removing most of the requests recorded in December and reducing the size of the web log by 6.7%. The effects of removing these requests can be seen in *Figure 6-10*.

Dataset 6 Nūpepa consists of the 2005 web log file created by the Ulukau web server, with most of the requests recorded in December removed. In total there were 286,383 requests to the Nūpepa digital library stored in a 286 megabyte file.

A comprehensive analysis of Dataset 6 Nūpepa was undertaken to determine how the Hawaiian language was used by clients of the Hawaiian Nūpepa digital library. This analysis is detailed in Chapter 6.

2.2.5 Summary of the Six Datasets of Web Logs

The six datasets of web logs are summarised in *Table 2-2*.

Summary of Web Log Datasets					
	Description	Web Server	Timeframe (all 2005)	Requests	Size (Mbyte)
1	NZDL	Puka, Thor	Jan-Dec	1,133,543	1,160
2	National	Puka	Jan-Dec	291,140	268
3	International	Thor	Jan-Dec	842,403	890
4	Default reo Māori	Puka, Thor	Jan, Mar, May, Jul, Sep, Nov	598,260	608
5	Default English	Puka, Thor	Feb, Jun, Aug, Oct, Dec	407,145	417
6	Hawaiian Nūpepa	Ulukau	Jan-Dec	286,383	268

Table 2-2: Information about web log datasets listing descriptions, web server, when recorded, requests and dataset file size

2.3 Preparing Datasets

As mentioned earlier, undertaking Transaction Log Analysis (TLA) involves four discrete steps: collecting the datasets, preparing the datasets, defining sessions, and undertaking analysis on the sessions.

Preparing the datasets involves removing unwanted data. This reduces processing time and ensures that the results are not influenced by unrelated data. Two types of data were removed from the datasets: redundant CGI (Common Gateway Interface) arguments and extraneous requests. The resultant datasets only contained data relevant to analysing how indigenous languages are used by clients of digital libraries.

2.3.1 Removing Redundant CGI Arguments

Each time a request is received by a web server, details about that request are recorded into a log file. The request details that are recorded vary depending on the type of web server and how the digital library has been configured. The Greenstone digital library software is usually configured to capture a large amount of data with each request. Below is an example of the entire request data recorded for a single request, with colour added to assist with explanation:

```

/cgi-bin/library 218-101-90-11.dialup.clear.net.nz [Thu
Jun 09 10:09:25 +1200 2005] (a=p, b=0, b1=0, b2=0,
bclaboutdesc=, bclcfgchanged=0, bclclone=0,
bclclonechanged=0, bclclonecol=, bclcontactemail=,
bcldirname=, bcldelete=0, bcleconf=0, bclesrce=0,
bclfromsrce=0, bclfullname=, bclinfochanged=0,
bclinput=, bclinputnum=3, bclinputtype=, bcltmp=, bcp=,
beu=, bft=, bnu=, bp=, bt=0, c=niupepa, cc=, ccp=0,
ccs=0, cfgfile=, cl=, cm=, cq2=, ct=0, d=, de=, debc=0,
dm=, ds=, dsbc=0, e=, el=prompt, er=, f=0, fc=1, fqa=0,
fqc=, fqf=, fqk=, fqn=4, fqs=, fqv=, g=, gc=0, gt=0,
h=, h2=, hd=0, hl=1, hp=, hs=0, ifl=, il=1, j=, j2=,
k=1, ky=, l=mi, m=50, n=, n2=, nl=, nw=utf-8, o=20,
p=about, pc=, pfd=0, pfe=0, pfl=0, pld=10, ple=10,
pll=10, ppnum=0, pptext=, pw=, pxml=0, q=, q2=, qb=0,
qf=0, qt=0, qto=3, r=1, rd=0, s=0, st=1, t=1, u=0, ua=,
uan=, ug=, uma=listusers, umc=, umnpw1=, umnpw2=,
umpw=, umug=, umun=, umus=, un=, us=invalid, v=0,
w=utf-8, x=0, xx=0, z=218.101.86.158-1118228138)
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"

```

This set of data consists of the following sections:

`/cgi-bin/library` - the directory on the web server where all the CGI (Common Gateway Interface) programs are stored. The CGI programs

allow data to be passed between host serving software and the client browsing software.

`218-101-90-11.dialup.clear.net.nz` —the IP Address and/or host name of the client browser that submitted the request.

`[Thu Jun 09 10:09:25 +1200 2005]` —the time stamp indicating when the request was received (Greenwich Mean Time is used).

```
(a=p, b=0, b1=0, b2=0, bclaboutdesc=, bclcfgchanged=0,
bclclone=0, bclclonechanged=0, bclclonocol=,
bclcontactemail=, bcldirname=, bcldelete=0,
bcleconf=0, bclesrce=0, bclfromsrce=0, bclfullname=,
bclinfochanged=0, bclinput=, bclinputnum=3,
bclinputtype=, bcltmp=, bcp=, beu=, bft=, bnu=, bp=,
bt=0, c=niupepa, cc=, ccp=0, ccs=0, cfgfile=, cl=, cm=,
cq2=, ct=0, d=, de=, debc=0, dm=, ds=, dsbc=0, e=,
el=prompt, er=, f=0, fc=1, fqa=0, fqc=, fqf=, fqk=,
fqn=4, fqs=, fqv=, g=, gc=0, gt=0, h=, h2=, hd=0, hl=1,
hp=, hs=0, ifl=, il=1, j=, j2=, k=1, ky=, l=mi, m=50,
n=, n2=, nl=, nw=utf-8, o=20, p=about, pc=, pfd=0,
pfe=0, pfl=0, pld=10, ple=10, pll=10, ppnum=0, pptext=,
pw=, pxml=0, q=, q2=, qb=0, qf=0, qt=0, qto=3, r=1,
rd=0, s=0, st=1, t=1, u=0, ua=, uan=, ug=,
uma=listusers, umc=, umnpw1=, umnpw2=, umpw=, umug=,
umun=, umus=, un=, us=invalid, v=0, w=utf-8, x=0, xx=0,
z=218.101.86.158-1118228138)
```

—the CGI arguments; settings used by the CGI programs to assist with information transfer. Many of these arguments are not used.

`"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"`

—the name and compatibility of the client's web browser (this is called the user agent string).

The CGI arguments from the web logs that are used and are relevant to this research are briefly described as:

<code>a=p</code>	action; p=page, q=query, d=document
<code>c=niupepa</code>	collection; which collection the request is associated with
<code>d=</code>	document identification (only set if a=d)
<code>g=</code>	document type (text, preview image, full image)
<code>k=1</code>	case-matching on queries (1=no matching, 2=case matching)
<code>l=mi</code>	language setting of interface (mi=reo Māori, en=English)
<code>m=50</code>	maximum results for query
<code>p=about</code>	page type (used if a=p); help=help page, about=about page
<code>q=</code>	query string; query term(s) submitted in a query
<code>s=0</code>	stemming (0=stemming off, 1=stemming on)
<code>t=1</code>	search type; some (t=1) or all (t=0)
<code>w=utf-8</code>	character set encoding
<code>z=218.101.86.158-1118228138</code>	unique user identifier (see §2.4.1)

A considerable amount of data is stored for each request in the log file. Almost three quarters of this data is redundant or not relevant to this particular research. Consequently, the first step in the preparation of the web logs was to remove all the CGI arguments that were not relevant to the analysis on indigenous language research. A text processing script, written in the Perl programming language, a Perl script, was created to remove the redundant data recorded for each request (see §2.5 for a summary of Perl scripts). This Perl script, called *process_raw_logs.pl*, was run on the web log datasets. Using the request data example listed above, the requests data remaining after *process_raw_logs.pl* was run is:

```
/cgi-bin/library 218-101-90-11.dialup.clear.net.nz [Thu
Jun 09 10:09:25 +1200 2005] (a=p, b=0, cl=, d=, gg=,
gp=, h=, j=, k=1, l=mi, m=50, p=about, q=, r=1, s=0,
t=1, v=0, w=utf-8, z=218.101.86.158-1118228138)
"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1)"
```

The request data example is thus reduced from 1041 characters to 262 characters, a reduction of 75%. Reducing each individual request data reduces the overall size of the web log datasets by a similar ratio, reducing the storage and processing requirements without compromising any of the relevant data.

2.3.2 Removing Extraneous Requests

The second stage of preparing the datasets is the removal of extraneous requests. The extraneous requests include: requests to other (non-Niupepa) digital libraries located in the NZDL, requests from non-human activity, requests made where the interface language is not reo Māori or English, requests with incomplete or incorrect settings and local requests submitted by developers of the Niupepa digital library. More than 80% of the requests collected in Dataset 1 were classified as extraneous requests and were removed from this indigenous language usage analysis. A description of these removals is given below.

Removing Non-Niupepa requests

The web log files generated by the Greenstone software record requests made to all digital library collections available at the New Zealand Digital Library (NZDL) web site (www.nzdl.org). There were 67 individual digital libraries available at the NZDL web site in 2005. The only requests appropriate to this research are requests to the Niupepa Collection (www.nzdl.org/niupepa). The Perl script *process_raw_logs.pl* was modified to also remove requests from Dataset 1 that were not requests to the Niupepa web site (i.e. requests were removed that did not have the CGI argument setting of `c=niupepa`).

Dataset 1 initially comprised 1,133,543 total raw requests. However 929,178 (82%) were requests to non-Niupepa digital libraries and were subsequently removed from this dataset. This left 204,365 requests remaining in Dataset 1 available for further processing.

Removing Requests From Web Robots

This research is concerned with indigenous language usage of the digital libraries by real users (i.e. human activity). Steps were undertaken to remove all requests that could be identified as being requested from Web robots. A web robot is a program that automatically travels the Web accessing web pages, and links on those pages.ⁱⁱⁱ Web robots are also known as web crawlers, web spiders and web wanderers. Two methods were used to identify requests from web robots: filtering for previously identified web robots, and analysing access behaviour for unusual, non-human like activity.

Previously identified web robots are defined by user agent strings and IP Addresses.

One task of the Perl script *process_raw_logs.pl* was to check the user agent strings stored in the web logs against a list of previously identified web robot user agent strings. The list was obtained from www.pgts.com.au/downloads.data in a file called *robots_list.txt*. At the time of the analysis, this file listed 327 robot user agent strings. When the Perl script *process_raw_logs.pl* found a user agent string match with a listed robot user agent string it removed the matched request from the dataset.

A similar method was used to check IP Addresses for known robot IP Addresses. The *process_raw_logs.pl* script also included code to check the IP Addresses in the dataset against a list of previously identified robot IP Addresses. This list was obtained from www.iplists.com and at the time of the analysis, contained 4,426 IP Addresses that had been previously identified as undertaking web robot activity. When the Perl script *process_raw_logs.pl* found an IP Address match with a listed robot IP Address it removed the corresponding request from the dataset.

ⁱⁱⁱ For more information about web robots see www.robotstxt.org/wc/robots.html and for lists of known web robots see www.dnsstuff.com.

These methods of identifying web robots were applied to Dataset 1 and identified a total of 292 web robots submitting 16,081 requests to the Niupepa web site.

Analysing access behaviour to identify web robots often occurs at a later stage in the TLA process. Tan and Kumar (2002) suggest that this method can be used in isolation to identify web robots. However, for the purposes of this research it was used as a secondary methodology. Identifying web robots by access behaviour involved defining the requests into sessions, defining the sessions into session types and then examining the session types over set time frames.

In the single request sessions (for a definition of single request sessions see §2.4.2) it became quite obvious that there were four IP Addresses with an abnormally high count of unique identifiers in a two month time period towards the end of 2005. These IP Addresses were creating 150–300 unique identifiers per month when the average number of unique identifiers created for that session type was just two per month. These four IP Address were investigated on the web site www.dnsstuff.com and were shown to be web robots that had recently been created. They were added to the list of known web robots and then the script *process_raw_logs.pl* was rerun.

Subsequently the Perl script *process_raw_logs.pl* identified a total of 296 web robots submitting 17,177 requests to the Niupepa Log. These requests were removed from the dataset.

While every endeavour has been made to remove web robot requests from the dataset it can be argued that requests from web robots may still exist in the log file. Perhaps some of the requests are from web robots that have not yet been identified. However, these requests are believed to be small in number and not likely to affect the overall results. Also, it is unlikely that such web robots requests submit queries and access actual pages of the Niupepa Collection, which is the focus of this TLA into how indigenous languages are used by clients of digital libraries.

Removing Requests With Inappropriate Language Settings

The analysis for this research was only concerned with requests submitted when the interface language was set to either reo Māori or English. All other requests were discarded. Aside from these two settings, the client has the ability to set the language of the Niupepa interface to one of 38 other languages. In addition, if the client accessed the site by using bookmarks, and some of the CGI arguments were altered throughout the year with software upgrades, as happened on a number of occasions, then the values in some of the CGI arguments may have been stored incorrectly. A typical example of this is a null value or a 00 value being stored in the CGI `l Argument`. The script *process_raw_logs.pl* also included code to remove all requests where the `l Argument` was not set to either `mi` or `en`. For Dataset 1 there were a total of 1018 requests (0.09%) removed because of an inappropriate `l Argument` setting.

Removing Requests With Other Incorrect Settings

Aside from the removal of requests outlined in the previous paragraphs, the Perl script *process_raw_logs.pl* was also configured to remove requests from the dataset where other CGI arguments were stored incorrectly. A request was removed from the dataset if there was an incorrect setting in the CGI `c (collection) Argument` or an incorrect setting in the `IP Address Argument`. The IP Addresses were also examined to detect the IP Addresses of computers that belonged to staff involved with maintaining and upgrading the digital library or web site. If the IP Address of a request was identified as belonging to a Niupepa staff member's computer, the request was removed from the dataset. The activity under investigation was the use of the Niupepa digital library by clients, not support staff.

2.3.3 Extraneous Requests Summary

A list of the extraneous requests removed from the initial raw log file Dataset 1 is displayed in *Table 2-3*. After the unwanted requests were removed, the remaining requests became the prepared log file, Dataset 1 Niupepa 2005. This completes the second step of TLA, preparing the datasets. The prepared log file was then available for the third step of TLA, defining sessions.

	Requests Removed	Requests
Initial Log file (Dataset 1)		1,133,543
Non-Niupepa Collection Requests	929,178	
Web Robot Requests	17,177	
Inappropriate language Argument	1,018	
Incorrect Collection Argument	4	
Incorrect IP Address	457	
Local IP Address	708	
Prepared Log File (Dataset 1 Niupepa)		185,001

Table 2-3: *Number of requests removed from initial raw log file Dataset 1 to create the prepared log file.*

2.4 Defining Sessions

The next stage in the TLA process was to collate the prepared datasets of requests into sessions so that indigenous language usage could be examined. Collating the requests into sessions involves three procedures: defining unique identifiers, defining session types and defining language of session types. These three procedures are explained in the next three sub-sections.

2.4.1 Defining Unique Identifiers

In order to collate requests into sessions the requests must be associated with an individual client. Identifying individual clients raises a whole range of issues. It is virtually impossible from a web log to identify specific individuals as insufficient client-explicit data is stored. Also, if specific clients could be identified then their permission would need to be sought before any personal information could be retained about them. The solution is not to identify clients as specific individuals but rather to identify them as unique identities whose personal details remain unknown. To identify clients as unique identities, a unique identifier is created and recorded with all activity undertaken by that client.

Two common techniques are currently used to associate clients with unique identifiers. The first technique is to use as a unique identifier the IP Address of the computer submitting the request. The second technique involves creating a unique identifier for each client and storing it on the client's computer in the form of an HTTP cookie^{iv}. Both techniques were available for the TLA undertaken on the Niupepa digital library as both techniques are possible using the Greenstone software.

As described in §2.3.1, the Greenstone software records the IP Address of the client browser with each request in the web logs. This IP Address can be used as a unique client identifier.

The Greenstone server software also uses HTTP cookies (known simply as cookies) by first querying the client's browser for a unique identifier cookie stored from the Niupepa web server. If the client's browser is able to respond with the Niupepa unique identifier cookie then the value stored in this cookie is assigned to the `Z Argument` for all hits generated by that client/browser (for an example of the `Z Argument` see §2.3.1 above). If the client doesn't have a Niupepa unique identifier cookie, one is generated by concatenating the client's IP Address with

^{iv} An HTTP cookie is a small text file created by a web server and saved by a client's web browser on the client's computer. The text of the cookie is sent to the web server each time the client makes a request to that web server.

the current UNIX time stamp. This method ensures no two unique identifiers will be assigned the same value. From then on this value, stored by client's browser as a Niupepa unique identifier cookie in the client's personal account and assigned by the web serving software to the `Z Argument`, can be used to uniquely identify the client.

There are advantages and disadvantages to both the IP Address technique and the cookie/`Z Argument` technique of defining unique identifiers.

The IP Address is always readily available; it is always supplied with requests and hence, using an IP Address to identify users is the method currently favoured by most web log analysis software. However, while the IP Address is often computer specific, it is not user specific; consequently if multiple users share one computer in sequence then several quite distinct sessions by different users might be classified as one single session. Further, people connecting through a shared dialup connection may also have a common IP Address and again, multiple users' activities might be classified as one single session.

Using cookies and the `Z Argument` to define unique identifiers eliminates the issue of users connecting through a shared dialup and it eliminates the issue of users using the same computer, providing they are logged onto the computer as different users. However, this technique will not identify separate users when multiple users are logging onto a shared computer using the same username (for example, if two or more people use a public library computer to access the collection). Also, users who do not enable cookies will never appear in the session data because a new unique identifier will be generated for each new request. Users who clear their stored cookies regularly will have a new unique identifier generated each time the cookies are cleared and this action will again affect the conclusions drawn in the session analysis.

IP Address or Cookies/Z Argument to Define Unique Identifiers?

Which technique was used to define unique identifiers in TLA undertaken in this thesis? The decision was straightforward. It was decided to use the Cookies/`Z`

Argument technique as although there were 4% fewer requests defined as actual usage sessions, there was more confidence that multiple users were not being defined as a unique identity.

2.4.2 Defining Session Types

Collating a prepared set of numerous requests into sessions of activity by individual clients is the core element of TLA that allows usage analysis. The most common method for collating requests into sessions is a timeout method. Using this method, sessions are defined as a sequence of two or more requests originating from a single client where the time between requests is no longer than a set interval. This set interval is usually 30 minutes.

The timeout method was used in this research to define sessions; however, the time interval was set higher than normal at 60 minutes. The time interval was increased because Niupepa users might be expected to spend extended periods reading the large amount of text on some of the newspaper documents and because the gap between requests might be exaggerated by web caching (which could mask intermediate requests).

It was decided to separate all prepared requests into one of three types of sessions: single request sessions, exploratory sessions and usage sessions. This decision was made for two reasons. First, as well as allowing for actual usage activities to be examined, it also allowed for the examination of exploratory activities, giving insights into why separate user groups may explore the web site but then not go on to access content stored in the web site. The second reason for defining the requests into three session types was to add another filtering stage so that the requests that were eventually defined as usage session requests were as unequivocal as possible.

Single request sessions were defined as requests submitted by a client when there was no other request made by that same client within the 60 minute time interval. These requests are not identifiable sessions as such but rather a random smattering of requests by seemingly different clients. Single requests could occur because of a client seeing one page of the web site and then deciding not to view any more pages. Single requests could also occur when a client has disabled cookies on their web browser.

An exploratory session was defined as a sequence of requests submitted by a single client with no more than a 60 minute time period between requests, no queries were submitted and no collection documents were accessed. Exploratory sessions constituted clients who accessed interface home pages, the help pages and/or the preferences pages but did not request any actual newspaper documents.

A usage session was defined as a sequence of requests submitted by a single client with no more than a 60 minute time period between requests, and, queries are submitted and/or collection documents requested. Usage sessions represent definitive activity by clients who were actually using the digital library to access content.

2.4.3 Defining Session Language Types

In order to analyse how indigenous languages were used by clients of digital libraries sessions were further defined into session language types. Session types were defined into two groups—those sessions where the indigenous language was used to undertake the activity and those where the non-indigenous language was used. For Dataset 1 Niupepa 2005, the sessions were divided into those sessions where the interface language was set to te reo Māori and those sessions where the interface language was set to English.

Defining the session language types for single request sessions is straightforward. Each request has the `L Argument` set to `mi` or `en`, that is, the language of the interface set to reo Māori or English. These interface language settings can be easily counted and compared. However, with exploratory sessions and usage sessions it is not so easy to classify the session as being a reo Māori session or an English language session as the sessions often comprise a combination of both language settings. Some decisions were made about what percentage of requests in te reo Māori constituted defining a session as a reo Māori session and what percentage of requests in English constituted defining a session as an English language session.

Exploratory sessions were assigned to session language types depending on which language the exploratory session was *mostly* conducted in. The `L Argument` was counted for each request in each session. If the `L Argument` was set to `mi` for more than 50% of the requests in the session, then that session was classified as an exploratory session *mostly* conducted in the medium of te reo Māori. For simplicity, the phrase ‘exploratory session mostly conducted in the medium of te reo Māori’ has been abbreviated to an ‘exploratory reo Māori’ session (`expMi`). Conversely, if the `L Argument` was set to `en` for more than 50% of the requests in the session, then that session was classified as an exploratory session *mostly* conducted in the medium of English. This phase has been abbreviated to an ‘exploratory English’ session (`expEn`). On those occasions when the `L argument` was set to one language for half of the session and the other language for half of the session then the session was classified by the language setting that the final request of the session was set to.

For usage sessions, the language classification was a little more complicated. Earlier research (Keegan & Cunningham, 2005a) had indicated that there were in fact three language classifications for usage sessions of the Niupepa digital library: usage sessions mostly in te reo Māori (abbreviated to ‘usage reo Māori’ (`usgMi`)), usage sessions mostly in the English language (abbreviated to ‘usage English’ (`usgEn`)), and bilingual sessions, that is, sessions where the user interface was set to reo Māori and English for a significant amount of the session (abbreviated to ‘usage Bilingual’ (`usgBi`)). After much consideration it was felt

that the earlier research may have classified many short sessions as bilingual sessions when, in fact, they were most likely monolingual clients taking more than one request to switch the interface language. Consequently, a new set of criteria for defining session language types was created, the details of which follow.

A usage session was defined as **usgMi** if it met the following criteria:

- L Argument = mi in more than 50% of the requests
- L Argument = mi in exactly 50% of the requests and
L Argument = mi in final request
- Session not defined as usgBi session

A usage session was defined as **usgEn** if it met the following criteria:

- L Argument = en in more than 50% of the requests
- L Argument = en in exactly 50% of the requests and
L Argument = en in final request
- Session not defined as usgBi session

A session was classified as **usgBi** if it met *all* of the following criteria:

- session must contain 10 or more requests
- L Argument =mi in more than 20% of the requests
- L Argument =en in more than 20% of the requests
- must contain 3 or more switches of interface language

The final requirement of the usgBi session, that bilingual sessions must contain at least 3 switches of interface language, was used to ensure that this grouping of sessions contained users who were specifically switching the language of the interface to assist with information retrieval. It was designed to eliminate clients who logged into the web site, used the interface in one language and then tried it in the other language but, on realising that this language was not as accommodating, then reverted back to the original language setting.

2.4.4 Collating Sessions

Two Perl scripts were written to collate the requests into sessions. A Perl script called *create-sessions.pl* collated all the prepared requests into one of the three types of sessions: single request sessions, exploratory sessions and usage sessions. The sessions were first collated using IP Address as a unique identifier and then collated again using the `Cookie/z argument` as a unique identifier. Both methods of creating sessions were undertaken for comparison purposes only, sessions created using the `Cookie/z argument` method were the only sessions used in the subsequent TLA processes.

A Perl script called *split-sess-lang.pl* divided exploratory and usage sessions into language sessions. All exploratory sessions were divided into one of two language session types; exploratory reo Māori sessions (`expMi`) and exploratory English sessions (`expEn`). All usage sessions were divided into one of three language sessions types, usage reo Māori sessions (`usgMi`), usage English sessions (`usgEn`) and usage bilingual sessions (`usgBi`).

2.5 Perl Scripts and Statistical Definitions

Text processing scripts, written in the Perl programming language were used to prepare the raw datasets of requests and collate these requests into appropriate session types for the analysis of indigenous language usage. Three Perl scripts have already been mentioned. The Perl script *process_raw_logs.pl* was written to remove redundant CGI arguments and extraneous requests. The Perl scripts *create-sessions.pl* and *split-sess-lang.pl* were written to define session types and session language types. Seven additional Perl scripts were written to generate a wide range of statistical data. The Perl scripts were written by Gareth Judsen under supervision in 2005/2006 while Gareth was a student of the Computer

Science Department at the University of Waikato. A summary of the ten Perl scripts is given in *Appendix A Perl Scripts Summary*.

A primary function of the Perl scripts was to generate statistical data to determine how indigenous languages are used by clients of digital libraries. This statistical data was output in text files which were transferred into spreadsheets. Spreadsheets were used to assist with comparative analysis and to generate graphs and tables. A listing of the statistical data generated by the Perl scripts is given in *Appendix A Perl Scripts Summary*.

The Perl Scripts were also used to generate indigenous language statistics on apparent reading and language switching by clients of a digital library. Apparent reading and language switching are explained in the next two sub-sections, followed by a sub-section explaining how statistical significance is defined.

2.5.1 Apparent Reading

The apparent read time is the length of time between two requests in a session. It is coined an apparent read time because it is impossible to confirm from the web logs that the user only spent their time reading that particular page or document before requesting the next page or document. With the large number of requests for the web interface pages and the Niupepa collection documents, and as we are only using these times to make comparisons against other times calculated in a similar manner, it is proposed that apparent read times can provide useful insights into indigenous language activity in a digital library.

There are two types of apparent read times—apparent page reads and apparent document reads. An apparent page read is the length of time between when a client requests a page of the web interface and when the client selects some other page or document. Apparent page reads are the lengths of time a client spends interacting with the web site interface. An apparent document read is the length of time between when a client requests a collection document, for example a newspaper page, and when the client selects some other page or document.

Apparent documents reads are the lengths of time that users spend reading the material of the collection.

Statistical results of apparent read times are discussed with other TLA statistical results in Chapters 3–6.

2.5.2 Language Switching Data

A language switch is defined as the changing of the interface language from reo Māori to English or from English to reo Māori. For the purpose of identifying the point at which the language switching was occurring, the sessions were divided into equal tenths. The point at which the language switch occurred was associated with the closest $1/10^{\text{th}}$ position in the session. The Perl script called *split-sess-lang.pl* was modified to generate language switching data. The language switching data included the number of language switches for each usage session type and the position of language switch within the usage session. Statistical results of interface language switching are discussed with other TLA statistical results in Chapters 3–6.

2.5.3 Determining Statistical Significance

Statistical significance is used in this research to decide whether differences observed between indigenous and non-indigenous users were real or whether the differences could have been caused by chance. Differences are observed in a wide range of areas and then mathematics is used to determine the probability that these differences are not caused by chance. If the probability (of chance) is low, then it can be assumed that the differences are real and are caused by some other factors. If the probability (of chance) is high then it can be assumed that there is statistically no significance in the difference.

The probability factor is referred to as a p value. Mathematical probabilities, like p values, range from 0 to 1, where 0 indicates there is no probability (that the difference is because of chance) and 1 indicates certainty (that the difference is because of chance). Generally speaking, if the p value is calculated to be less than 0.05 then the difference is said to be statistically significant. However, to give even more certainty, a p value lower than 0.01 will be used in this thesis to indicate differences are statistically significant.

The chi square test can also be used to determine the significant difference in data. It is particularly applicable for use in results from TLA as it can be easily applied to test for significance in bivariate tables. The statistics produced from the TLA in this research is often presented in bivariate tables. Consequently, p values, generated from chi square tests, are used to determine if the differences in these values are statistically significant.

2.6 Summary

This chapter has explained in detail the Transaction Log Analysis methodology used to define how indigenous languages are used by clients of digital libraries. The collection of the six datasets of web log requests used in this research has been described. The methods used to prepare those requests and filter out redundant CGI arguments and extraneous requests have been described. The definitions used to collate the resulting requests into session types and language of session types have also been explained. The Perl scripts used to generate statistical information from the language session types have also been summarised. The next stage is to examine the statistical data generated by the TLA process to determine how indigenous languages were used by clients of digital libraries.

Statistical data generated about indigenous language usage of the Niupepa Collection is analysed in Chapter 3. Statistical data analysing the difference between national and international indigenous language usage of the Niupepa

collection is revealed in Chapter 4. Statistical data interpreting the influence of the default language setting on indigenous language usage is examined in Chapter 5. Statistical data generated about indigenous language usage of the Hawaiian Nūpepa Collection is analysed in Chapter 6. These four chapters comprehensively examine a very large amount of data on the issue of how indigenous languages were used by clients of digital libraries.

3 Reo Māori Usage of the Niupepa Digital Library

Chapter 3 discusses how te reo Māori was used by clients of the Niupepa digital library in 2005.

A prime question that must be asked before the discussions begin is, was te reo Māori used by Niupepa clients? Results from Transaction Log Analysis (TLA) indicated that te reo Māori was used significantly by clients in single request sessions, exploratory sessions and usage sessions. Having established that te reo Māori was used, the next stage is to examine how it was used by Niupepa clients.

Data from exploratory sessions gave insights into how clients explored the Niupepa interface in te reo Māori but then did not go on to use the Niupepa digital library. This data showed that Niupepa exploratory sessions in te reo Māori (expMi) were short, with the Niupepa home page being the interface page that received the most requests.

Data from the usage sessions was then examined to determine how reo Māori clients used the Niupepa interface, how reo Māori clients accessed documents and how reo Māori clients submitted queries. These three topics give a comprehensive account of how te reo Māori was used by Niupepa clients.

Statistics generated by TLA were also used to compare reo Māori/English language usage and examine bilingual use. The reo Māori/English language comparison reveals different session lengths, different reading times and different querying behaviours between usgMi clients and usgEn clients. The bilingual

usage (usgBi) analysis reveals usgBi sessions to be long, with a focus on using queries to access documents.

This chapter begins by presenting a description of the interface screens and options that are available to clients of the Niupepa digital library as context for discussing reo Māori usage.

3.1 Niupepa Interface Description

The Niupepa digital library has a multilingual interface that serves a collection of legacy newspapers written mostly in te reo Māori. The Niupepa digital library can be found at www.nzdl.org/niupepa. The web site offers different methods for accessing the newspaper documents and the documents themselves are available in various formats. As well as newspaper documents, additional information about the newspapers is available from the Niupepa web site. The methods available for accessing the Niupepa documents, the document formats and additional documents available to the Niupepa clients are displayed in the screenshots below.

The home page of the Niupepa web site is displayed in *Figure 3-1*. The interface is displayed with the default language set to reo Māori; the usual default language setting for this web site. The language of the interface can be easily changed to English by clicking on the English hyperlink which is located towards the top of the home page. The home page is the only web site page on which the language of the interface can be directly alternated between English and reo Māori. Elsewhere the interface language can be switched by using the Preferences page (as displayed in *Figure 3-2*).



Figure 3-1: The home page of the Niupepa web site (reo Māori interface).

The Niupepa home page is divided into three sections. Along the top right of the page are the Āwhina (Help), Kōwhiringa (Preferences) and NZDL buttons which link to their respective web pages. Below these buttons is the title of the page (kāinga or home) and then the access bar. The access bar contains three primary methods to access the documents. The Rapu (Search) button invokes the search engine shown in *Figure 3-3*. The Tānga Pukapuka (Series) button will display a series list of all the newspapers as shown in *Figure 3-4*. The Ngā Rā (Dates) button will show a list of newspaper documents published in a particular time period, as shown in *Figure 3-5*. The home page text gives background information about the Niupepa collection.

The Preferences (Kōwhiringa) page of the Niupepa digital library web site is displayed in *Figure 3-2*.

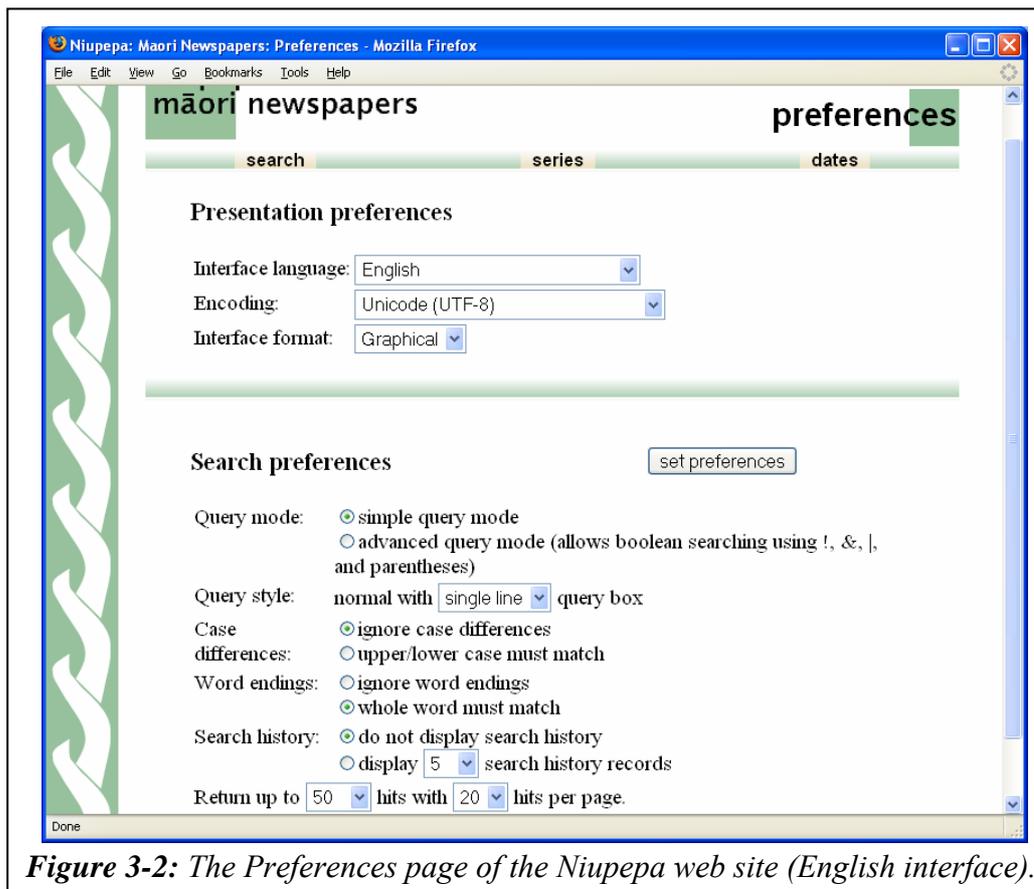


Figure 3-2: The Preferences page of the Niupepa web site (English interface).

Accessing Niupepa Documents

There are three methods by which clients may access Niupepa documents. Clients can submit a query as displayed in *Figure 3-3*. This example shows the results page following the submission of a query for ‘Paratene Ngata’ⁱ. The second method is to select a periodical from the series list, as displayed in *Figure 3-4*. Clicking a periodical title from this page will display a list of issues available from that newspaper. The third method of accessing Niupepa documents is by using the Date facility, as displayed in *Figure 3-5*. This example shows all the available documents published in the year 1900.

ⁱ Paratene Ngata is an ancestor of the author from the Ngāti Porou tribe.

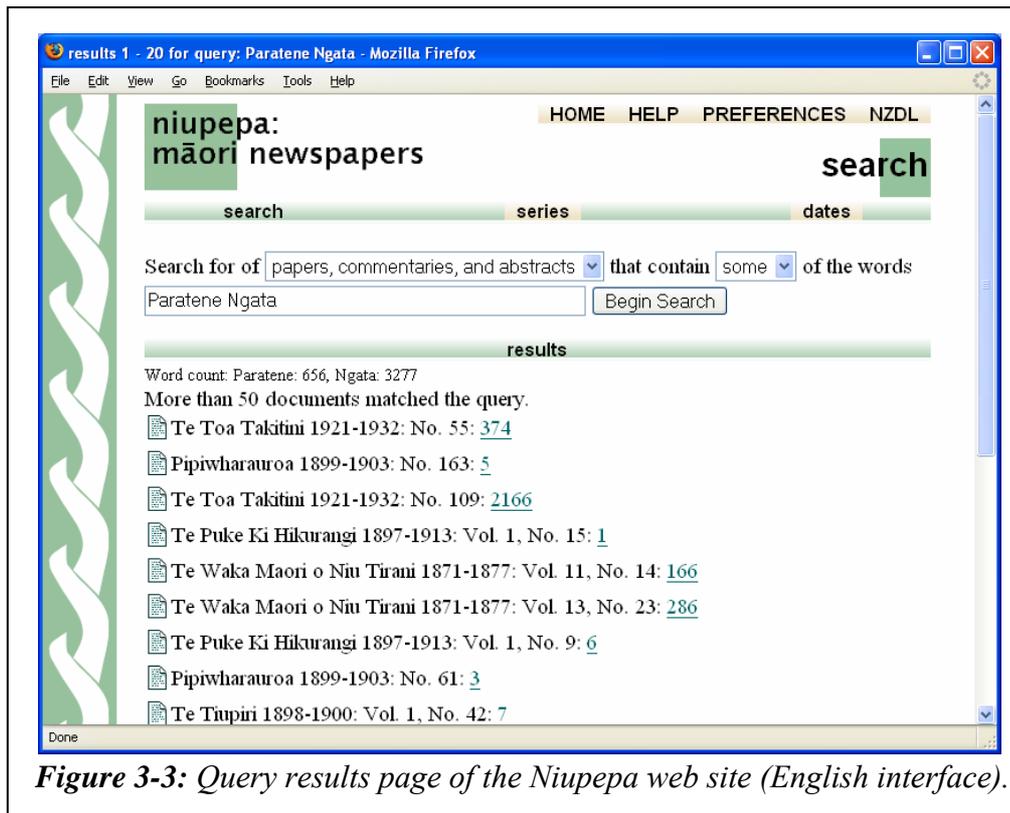


Figure 3-3: Query results page of the Niupepa web site (English interface).



Figure 3-4: Series list of the Niupepa web site (English interface).

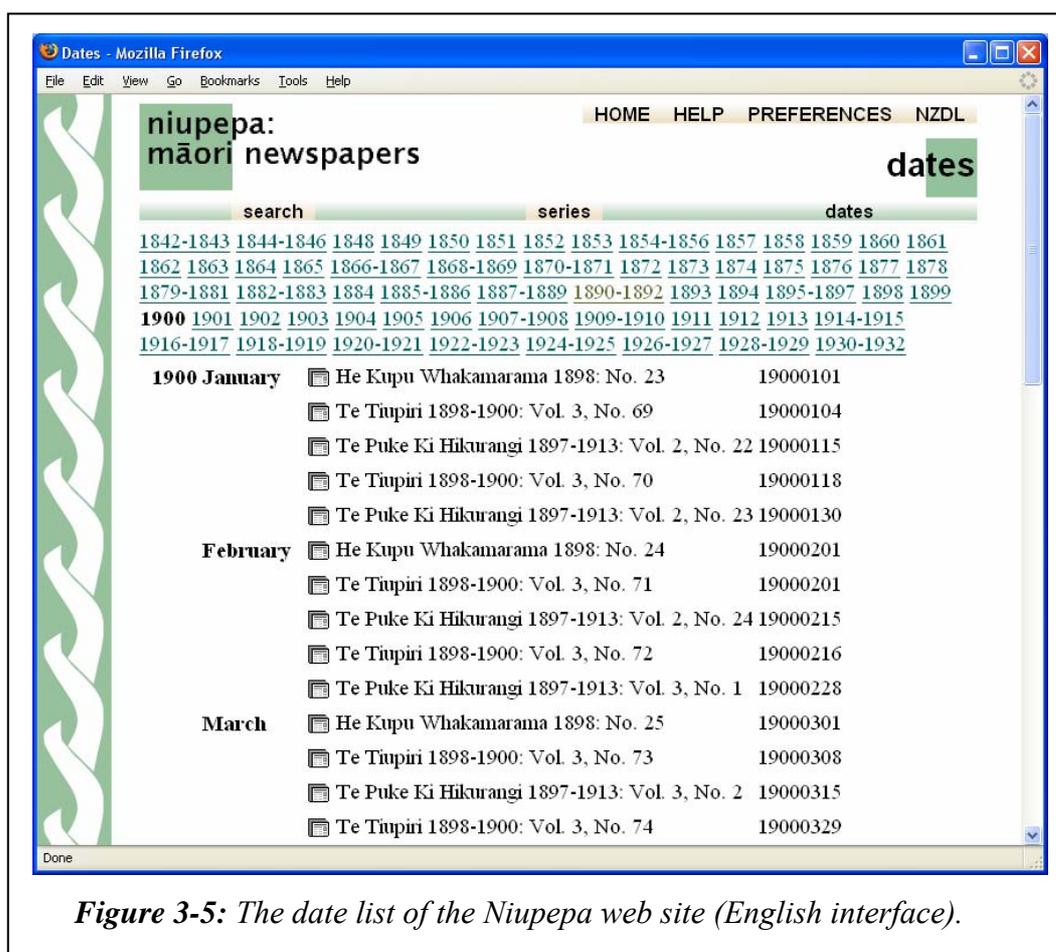


Figure 3-5: The date list of the Niupepa web site (English interface).

Figure 3-3 also shows the two query parameters that the client can alter on the Query page. The first parameter is papers, commentaries and abstracts. By default the texts of the newspapers, the bibliographic commentaries and the English abstracts are searched collectively but an option is available to search separately on any one of these resource types. The bibliographic commentaries and the English abstracts are described below. The second query parameter is the Some/All parameter which is used when more than one query term is entered. It gives the user the option to seek documents that contain *some* of the query terms or documents that contain *all* of the query terms.

Niupepa Document Formats

The Niupepa web site displays the Niupepa documents in three formats, extracted text, preview image and full size image. An example of the extracted text format is shown in *Figure 3-6*, an example of the preview image format is shown in *Figure 3-7*, and an example of the full size image format is shown in *Figure 3-8*. The examples listed show a document that has been requested following a query for the terms ‘Paratene Ngata’. Highlighting of the query terms only occurs in the extracted text version of the document.

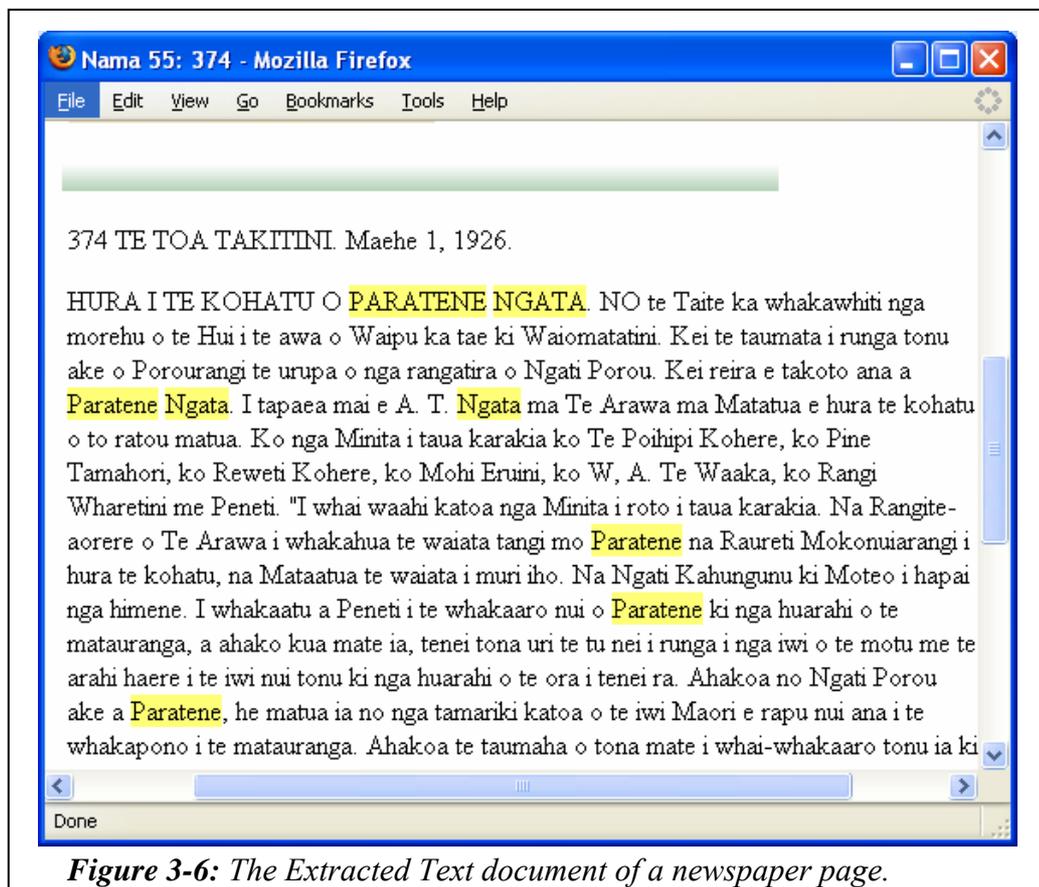


Figure 3-6: The Extracted Text document of a newspaper page.

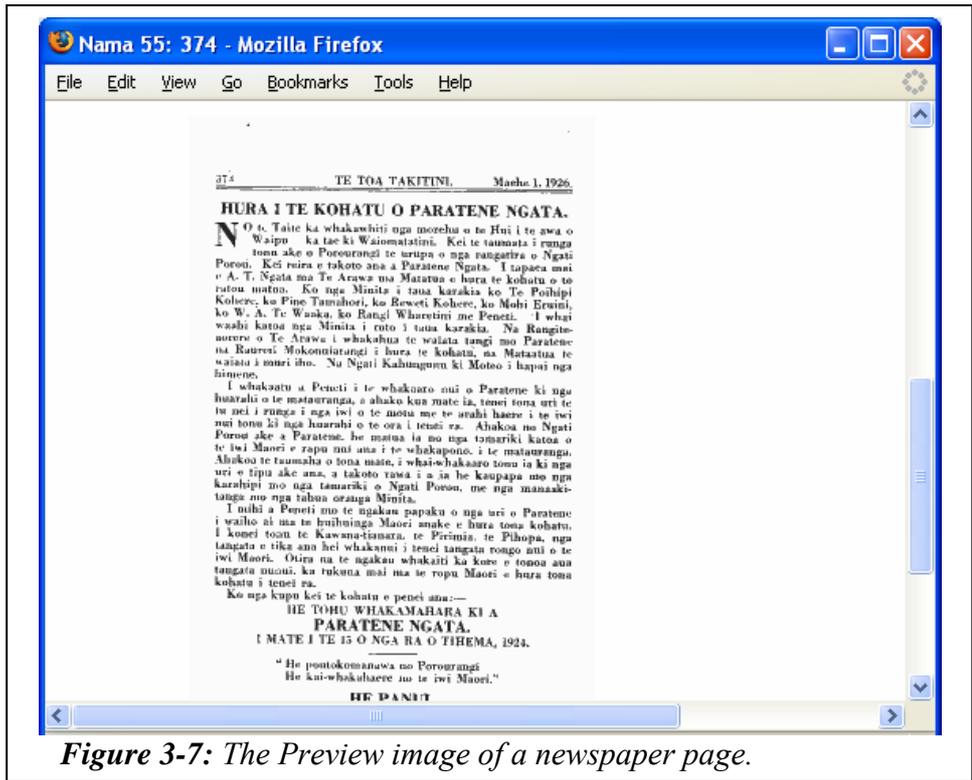


Figure 3-7: The Preview image of a newspaper page.

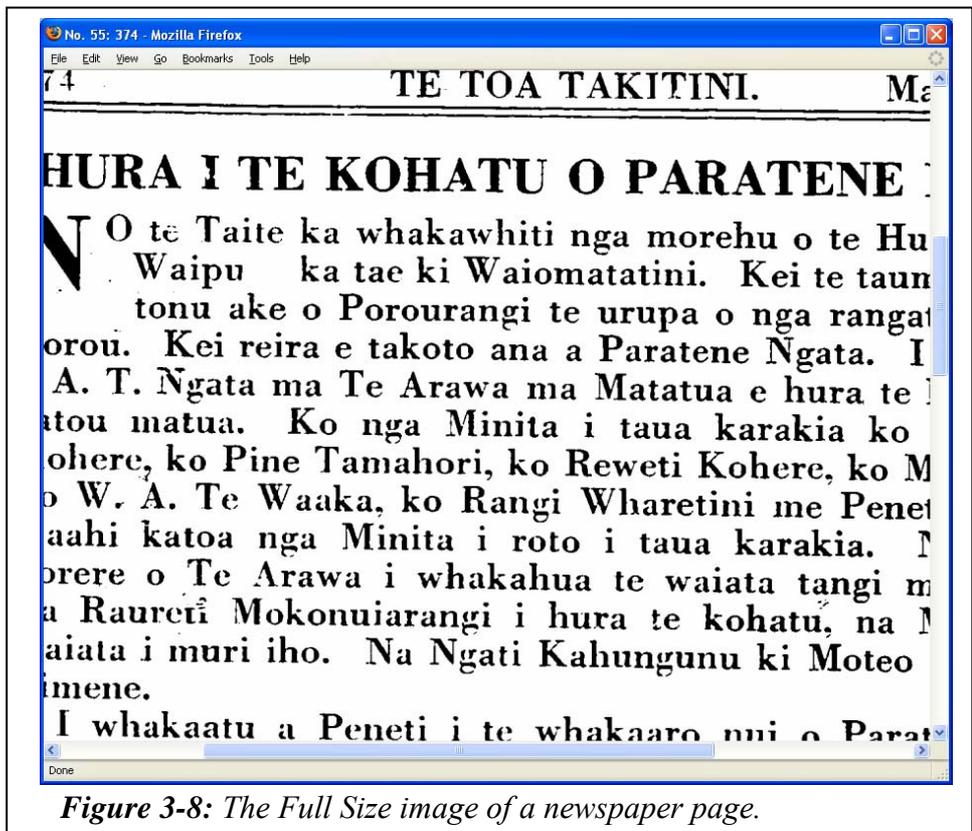


Figure 3-8: The Full Size image of a newspaper page.

The Niupepa web site also makes available two additional types of information: English abstracts and bibliographic commentaries. The English abstracts were compiled by a team of researchers at the University of Auckland headed by Dr Jane McRae. The abstracts are summaries (in English) of over 40% of the reo Māori newspapers and give non-Māori literate users an understanding of what has been written. An example of an English abstract is shown in *Figure 3-9*.

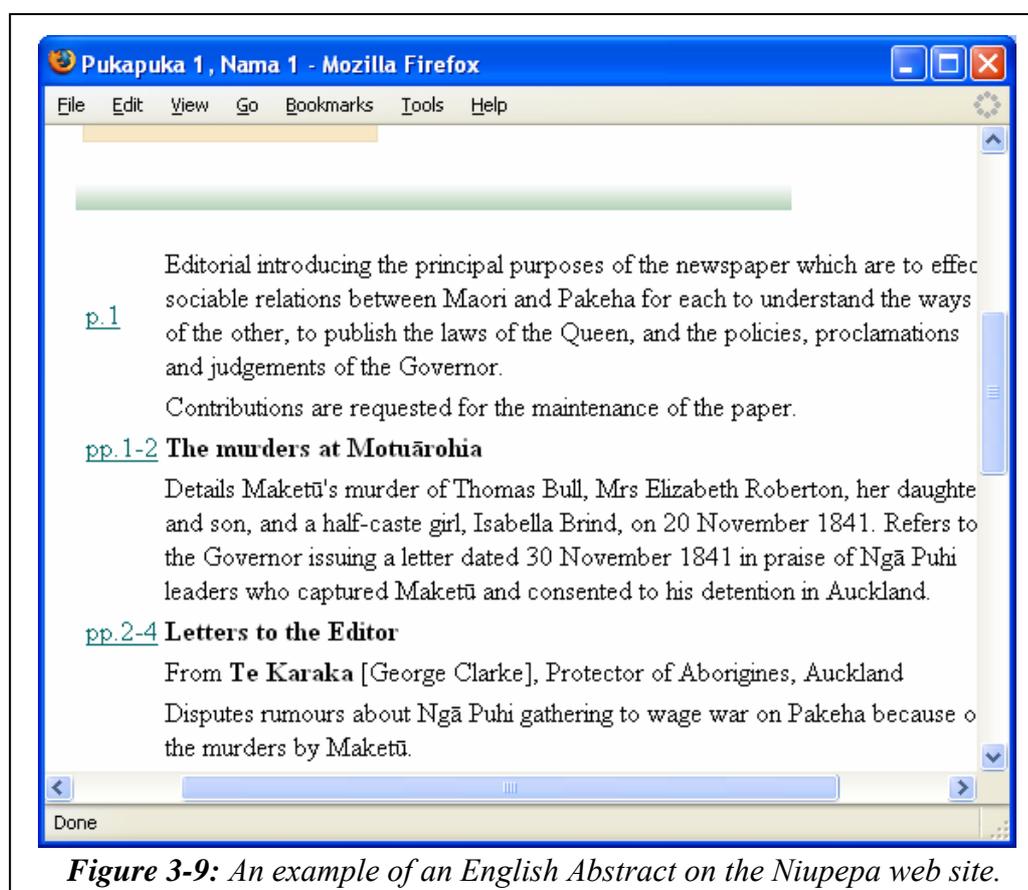


Figure 3-9: An example of an English Abstract on the Niupepa web site.

The bibliographic commentaries were written by Gail Dallimore (Dallimore, 1990) and give background information about each periodical, including publication details, subject matter and where surviving originals can be sourced. An example of a Bibliographic Commentary is shown in *Figure 3-10*.

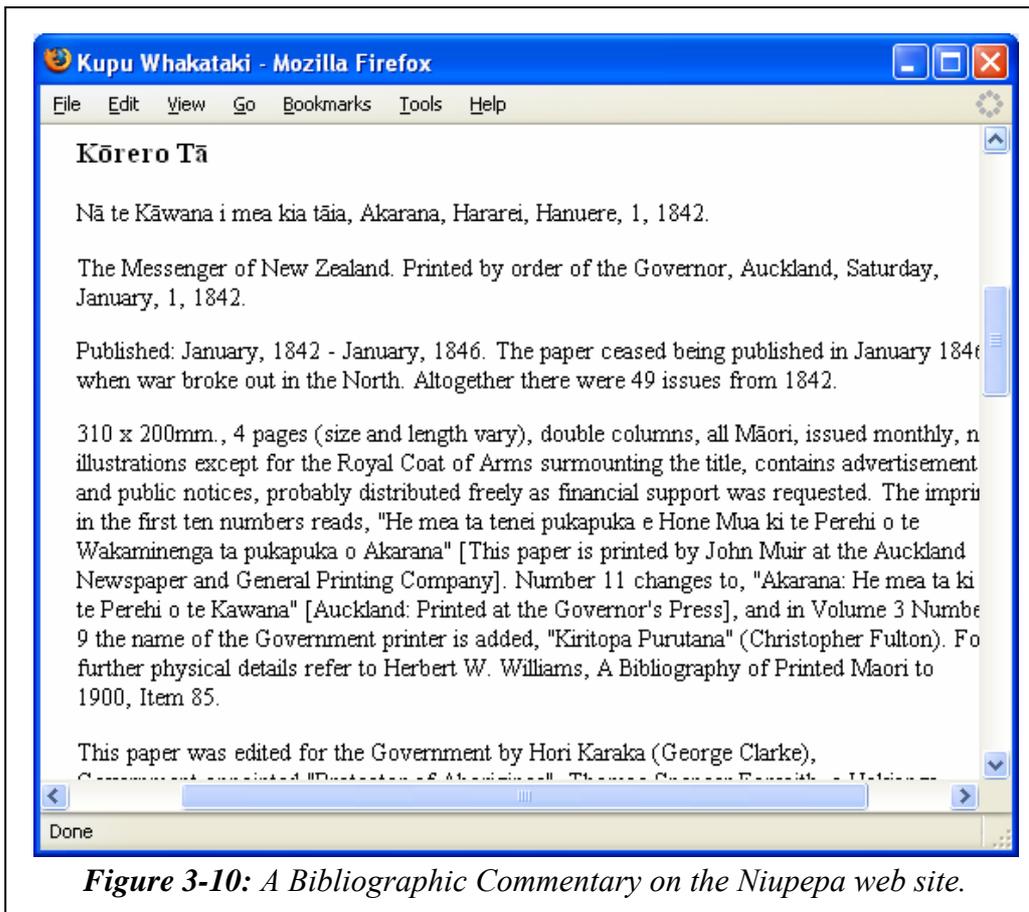


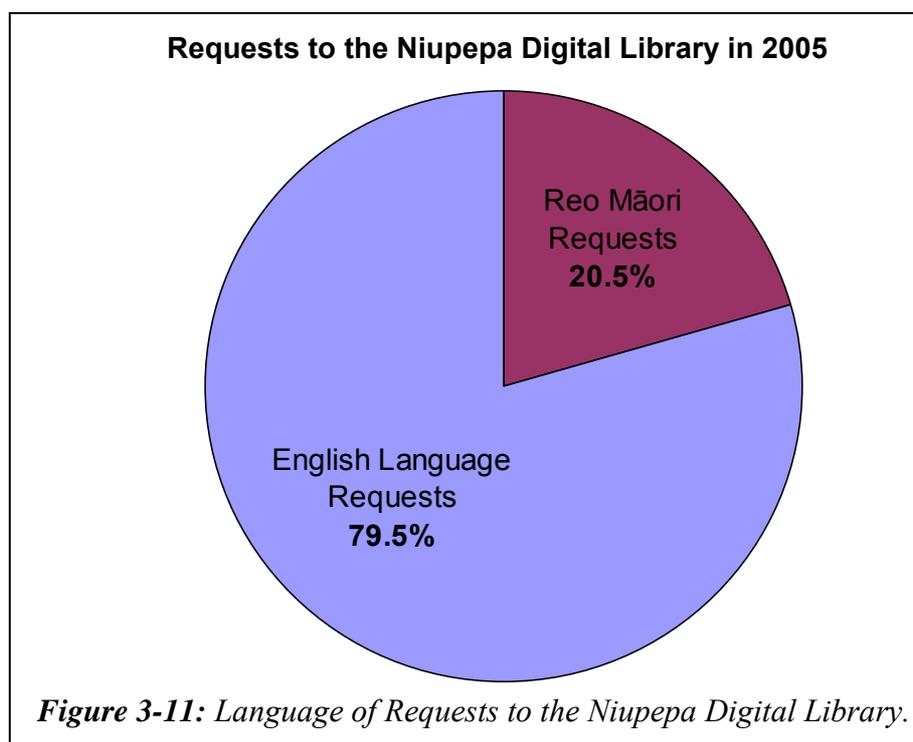
Figure 3-10: A Bibliographic Commentary on the Niupepa web site.

Figure 3-9 and *Figure 3-10* bring to a close the description of the Niupepa interface. The following section of this chapter will begin to examine how this interface was used in te reo Māori by clients of the Niupepa digital library.

3.2 Was Te Reo Māori Used by Niupepa Clients?

The first question to be answered from the TLA statistics is, was te reo Māori used by clients of the Niupepa digital library?

The Niupepa web site received 1,133,543 raw requests in 2005. The first request was logged at 12:02am January 1st 2005 and the last request for 2005 was logged at 11:58pm December 31st 2005. This raw log of requests was filtered to remove all the unwanted requests (as described in §2.3.2) leaving a filtered total of 185,001 requests (see *Table 2-3*). Of this total, 37,936 (20.5%) were requests with the interface language set to reo Māori and 147,065 (79.5 %) were requests with the interface language set to English. These requests are displayed in *Figure 3-11*.

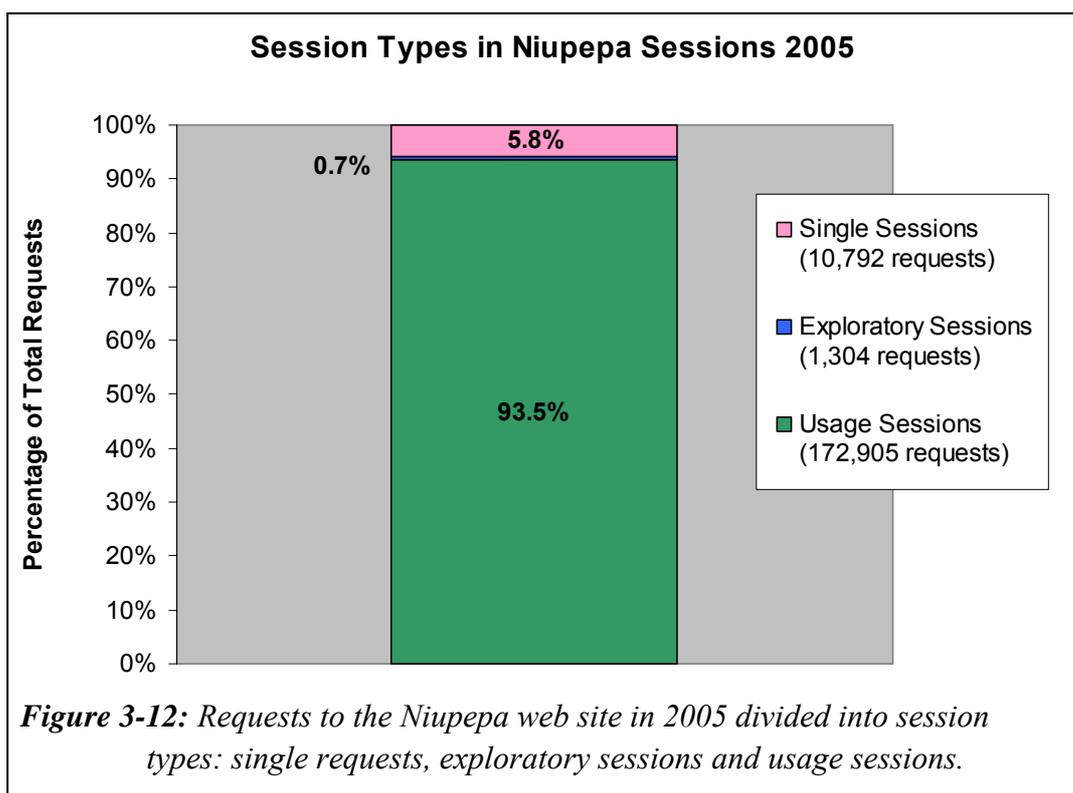


Consequently, we can state that te reo Māori was used in approximately 20% of client requests to the Niupepa digital library. However, was it used in all three of the session types?

3.2.1 Was Te Reo Māori Used in all Session Types?

The TLA process separated the requests into the three session types (as described in §2.4.2). There were 10,792 requests (5.8%) classified as single request sessions, 1,304 requests (0.7%) classified into 567 exploratory sessions, and 172,905 requests (93.5%) classified into 6,744 usage sessions. The percentages of requests in the different session types are displayed in *Figure 3-12*.

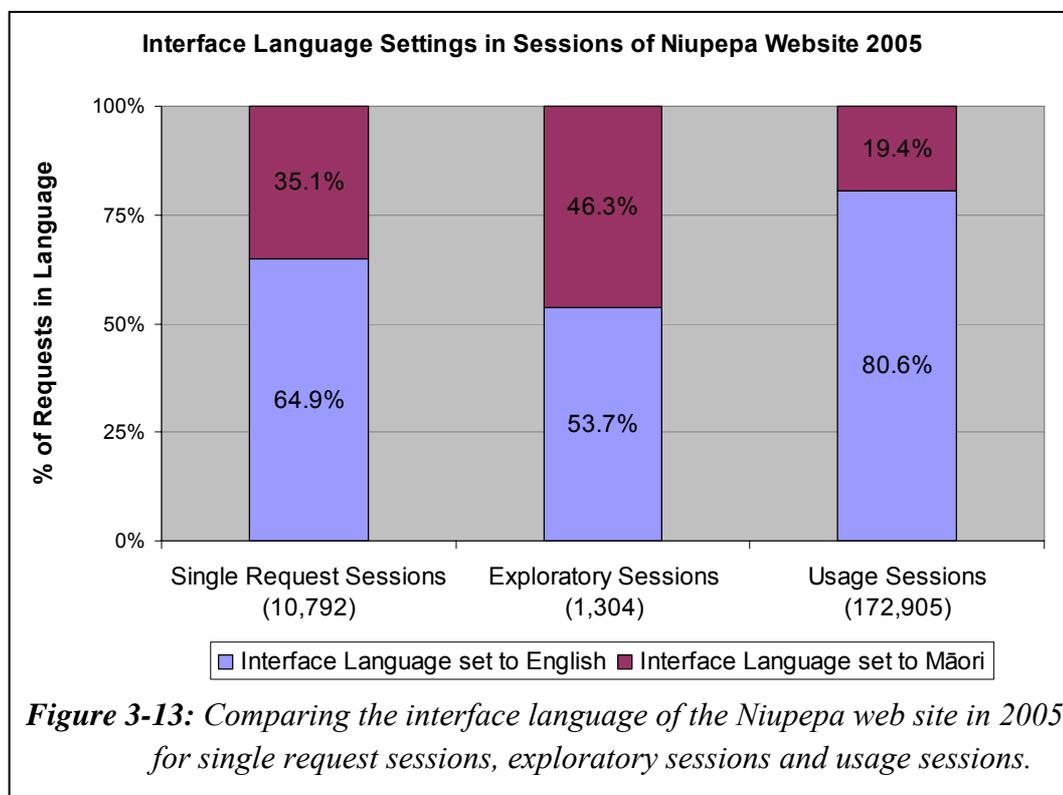
Figure 3-12 shows the majority of filtered requests were usage session requests, that is, requests from clients who were actively engaging with the Niupepa digital library to read newspaper documents and/or undertook a query to locate those documents.



The interface language settings from the TLA were examined for the three session types. For single request sessions there were 3,792 requests when the interface language was set to te reo Māori and exactly 7,000 when the interface language was set to English. For exploratory sessions, there were 604 requests when the

interface language was set to te reo Māori and 700 when the interface language was set to English. For usage sessions, there were 33,540 requests when the interface language was set to te reo Māori and 139,365 when the interface language was set to English. These figures have been calculated as percentages and are graphed in *Figure 3-13*.

Figure 3-13 clearly shows that te reo Māori was being used by clients of digital libraries in all three session types.



As the Niupepa web site's default language was set to reo Māori for approximately 60% of the time used in the web log data, one could suppose that there would be a 60/40 ratio of single requests in reo Māori and English. The 35/65 ratio suggests that some of the requests may have been incorrectly defined as single requests, perhaps caused by users disabling cookies or by network and browser caching which would prevent common requests being received by the Niupepa web server.

The Exploratory Sessions show an almost 50:50 ratio of reo Māori use, confirming that te reo Māori was being used by clients of the Niupepa digital library to explore the Niupepa collection.

The usage sessions show that te reo Māori was being used in almost 20% of the usage requests. This percentage indicates significant use and 33,540 usage requests in te reo Māori give sufficient data for further TLA to determine exactly how te reo Māori was used by clients of the Niupepa digital library.

3.2.2 When Was Te Reo Māori Used?

There are three timeframes to be considered in analysing when te reo Māori was used by clients of the Niupepa digital library. What time of the day was it used? What day of the week was it used? And what month of the year was it used?

Requests from exploratory sessions in te reo Māori (expMi) were combined with requests from usage sessions in te reo Māori (usgMi) and the percentages were plotted in *Figures 3-14, 3-15 and 3-16*. As a language comparison, the English exploratory and usage requests (expEn and usgEn) were also plotted in *Figures 3-14, 3-15 and 3-16*.

*Figure 3-14*ⁱⁱ shows the requests grouped per hour. Te reo Māori was mostly used between the hours of 9:00am and 10:00pm, with a clear peak between 1:00pm and 3:00pm. A similar usage pattern was noted when the digital library was being used in English; however, the English language requests did not peak to the same extent just after midday.

ⁱⁱ Note: this data is only taken from national requests where it is certain that the time zone of the client matched the time zone recorded in the Web log.

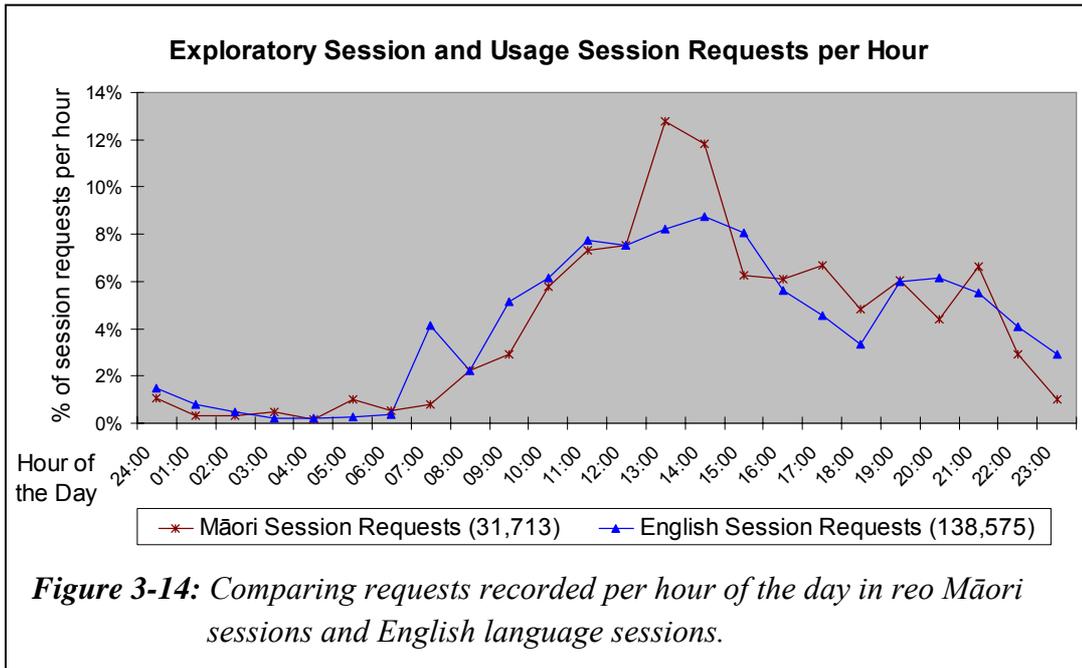
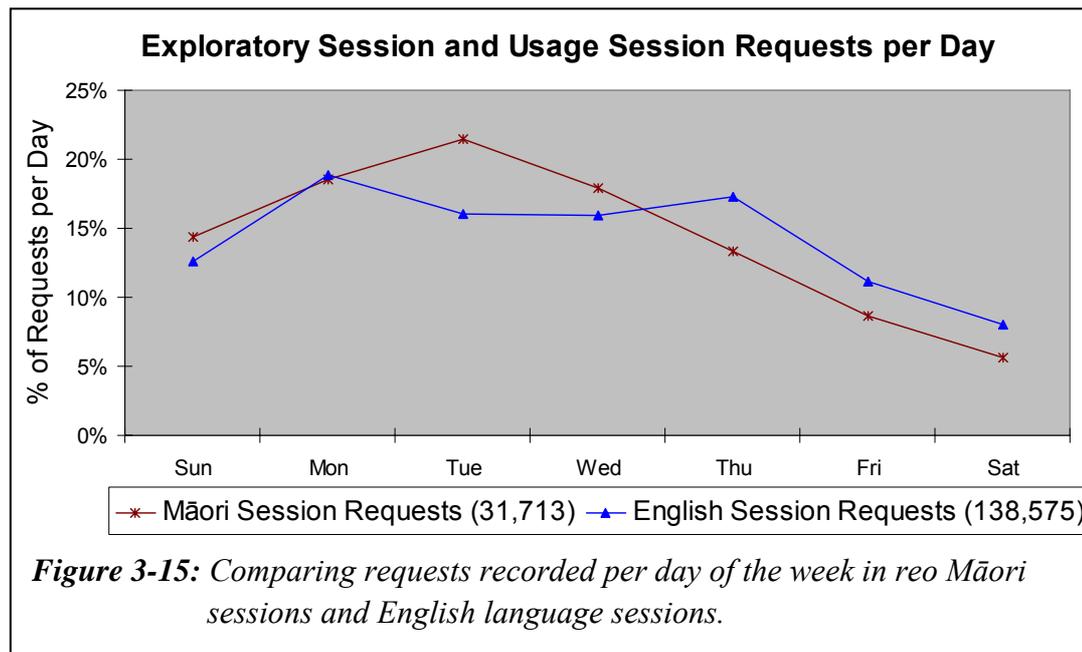
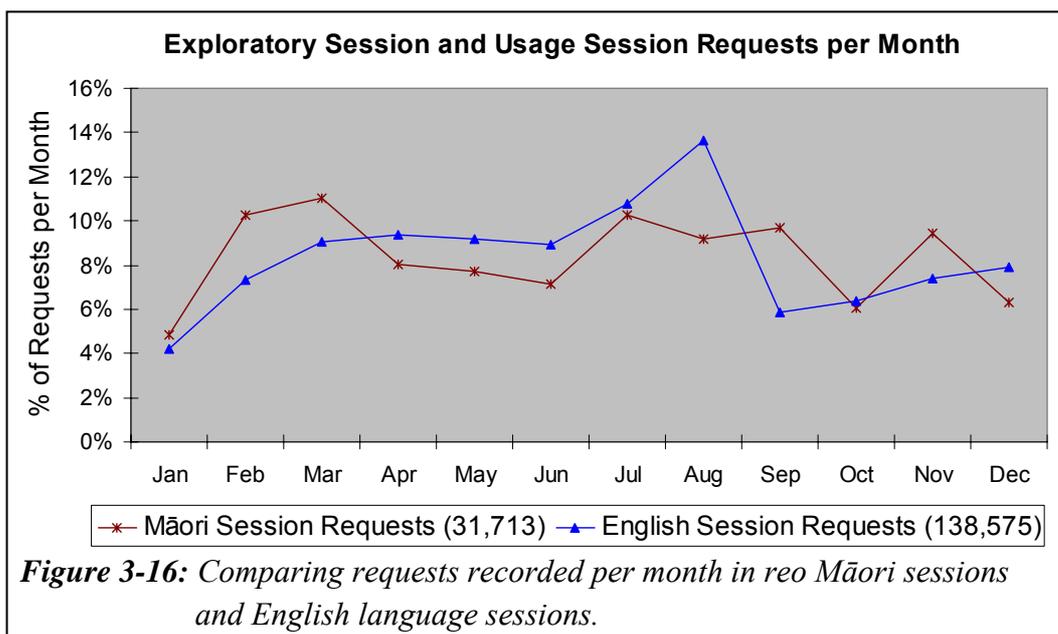


Figure 3-15ⁱⁱⁱ shows use of te reo Māori was mostly occurring on Mondays through to Wednesdays, with the least use occurring on Fridays and Saturdays. This usage was similar to the usage seen in the English language, although English language clients showed a preference to use the digital library on Mondays through to Thursdays.



ⁱⁱⁱ See previous note.

Figure 3-16 shows requests in te reo Māori grouped by month, as they occurred in 2005. Excluding January, the variance in months is never more than 3.2% indicating that te reo Māori was used consistently throughout the year. This consistency was not reflected in the English language usage, which showed a peak in August, followed by a trough in September.



In summary, te reo Māori was used by clients of the Niupepa digital library in 2005, mostly between 9:00am and 10:00pm with a peak between 1:00pm and 3:00pm, mostly on Mondays, Tuesdays and Wednesdays, and reasonably consistently throughout the year.

Now that it has been established that te reo Māori was being used and when it was being used, the next stage is to identify exactly how it was being used. The remainder of this chapter describes client behaviour and is the result of a comprehensive Transaction Log Analysis undertaken on the Niupepa web logs.

3.3 Exploratory Behaviours

Requests made by clients that explore the Niupepa digital library (but then do not go on to use it) represent less than 1% of the total filtered requests. While this is a low number of requests, it does represent 5.7% of the combined number of exploratory and usage sessions. This data can highlight how these types of clients, exploratory users, use the Niupepa interface before deciding not to access any documents or submit any requests. In particular, this section will examine exploratory sessions by users who explore in the medium of te reo Māori (expMi).

There were 1304 requests recorded in 567 exploratory sessions for the Niupepa collection in 2005. A high proportion, 84.0%, of the sessions were less than three minutes and a high proportion, 93.7%, consisted of three requests or less. Note that two requests was the minimum requirement for classification as a session.

The question we now ask is, how was te reo Māori used in exploratory sessions by clients of digital libraries? To answer this question the exploratory sessions were subdivided into the two language categories. There were 180 exploratory sessions where the Niupepa web site interface language was mostly set to reo Māori (expMi). In these sessions the interface language was set to reo Māori for 81.2% of the requests. There were 387 exploratory sessions where the Niupepa web site interface language was mostly set to English (expEn). In these sessions the interface language was set to English for 70.6% of the requests.

TLA was used to examine three characteristics of exploratory sessions in te reo Māori: the length of expMi sessions, the apparent reading times in expMi sessions and the interface pages requested in expMi sessions.

3.3.1 How Long Were Exploratory Sessions?

Session length is measured in two categories—duration in time (seconds) and the number of requests. The comparative lengths of expMi and expEn sessions in seconds are displayed in *Table 3-1*. The comparative lengths of expMi and expEn sessions by number of requests are displayed in *Table 3-2*.

expMi			expEn		
Seconds	Sessions	%	Seconds	Sessions	%
Total	180	100.0%	Total	387	100.0%
0–1	36	20.0%	0–1	38	9.8%
2–9	17	9.4%	2–9	85	22.0%
10–29	40	22.2%	10–29	116	30.0%
30–179	50	27.8%	30–179	94	24.3%
180+	37	20.6%	180+	54	14.0%

Table 3-1: Comparisons of length (in seconds) of expMi and expEn sessions.

expMi			expEn		
Requests	Sessions	%	Requests	Sessions	%
Total	180	100.0%	Total	387	100.0%
2	133	73.9%	2	329	85.0%
3	34	18.9%	3	35	9.0%
4	9	5.0%	4	11	2.8%
5–6	4	2.2%	5–8	12	3.1%

Table 3-2: Comparisons of number of requests in expMi and expEn sessions.

ExpMi sessions were short. In terms of time, 20% were less than two seconds and 51.6% were less than 30 seconds. In terms of requests, 73.9% of sessions were the bare minimum of two requests and 92.8% of sessions consisted of just two or three requests.

ExpMi sessions were compared with expEn sessions to see if the sessions were similar in length. Initial glances at both tables seem to indicate that the expMi and expEn sessions were similar in length, both in duration and number of requests. A chi square statistical analysis undertaken on the length of sessions in seconds gives a p value of $1.9 * 10^{-5}$ and in requests gives a p value of $3.5 * 10^{-3}$. These values indicate that there was a significant statistical difference in the duration (in time) between expMi and expEn sessions but not a statistical difference in the number of requests. These values also suggest that clients in expMi sessions spent a significantly longer time reading the interface pages of the Niupepa web site than clients in expEn sessions. This supposition can be confirmed by examining how long it took expMi clients and expEn clients to read interface pages of the Niupepa digital library.

3.3.2 How Long to Read Interface Pages?

The length of time from when a client requests a page of the interface to when the client requests another page of the interface, the apparent page read time, was calculated. Since it was not possible to confirm that the clients were only reading the pages of the interface during this time, these times have been labelled apparent page read times. These times give indications about how long clients take to read interface pages.

The apparent page read times were grouped in intervals of 1–2 seconds, 3–9 seconds, 10–19 seconds, 20–39 seconds, 40–79 seconds and 80+ seconds. The apparent page read times for both expMi and expEn sessions are displayed in *Table 3-3*.

Apparent Read Times in Exploratory Sessions

Seconds	expMi	expEn
Total Apparent Reads	246	491
1–2	55	59
3–9	36	101
10–19	33	89
20–39	37	88
40–79	21	57
80+	64	97

Table 3-3: expMi and expEn apparent read times of Niupepa interface pages in exploratory sessions.

A chi square analysis was undertaken on the apparent read times and returned a p value of $5.6 * 10^{-4}$. This value indicates that there was a significant statistical difference in the expMi and expEn apparent read times of interface pages. To determine exactly what this difference was, statistical summaries of the apparent read times were calculated and are displayed in *Table 3-4*.

Statistical Summary for Exploratory Apparent Read Times

	expMi	expEn
Total Apparent Reads	246	491
Average (seconds)	145.6	106.9
Standard Deviation (seconds)	410.1	339.7
Mode (seconds)	0	0
Shortest session (seconds)	0	0
Lower Quartile (seconds)	4	8
Median (seconds)	19	19
Upper Quartile (seconds)	74	53
Longest session (seconds)	3396	3497

Table 3-4: Statistical summaries comparing apparent read times in expMi and expEn exploratory sessions.

These statistics displayed in *Table 3-4* show that the average apparent read time of expMi sessions was almost 2.4 minutes (145.6 seconds). The average apparent read time of expEn sessions was 1.8 minutes (106.9 seconds). Both medians are equal at 19 seconds but the higher upper quartile of expMi suggests that these sessions had a higher proportion of longer exploratory sessions.

These statistics suggest that exploratory clients of the Niupepa digital library who used the interface in te reo Māori took more time to read the interface pages than clients who used the interface in the English language.

3.3.3 Which Interface Pages Were Requested?

The interface pages requested by expMi and expEn clients are listed in *Table 3-5*. Note that an interface page refers to a page of the Niupepa web site; it does not refer to a newspaper document in the Niupepa collection.

	expMi		expEn	
Total pages	426	100.0%	878	100.0%
Niupepa home page	358	84.0%	728	82.9%
NZDL home page	9	2.1%	60	6.8%
About English Abstracts	38	8.9%	35	4.0%
help page	5	1.2%	24	2.7%
preferences page	15	3.5%	29	3.3%
about greenstone page	0	0.0%	1	0.1%
unknown page	1	0.2%	1	0.1%

Table 3-5: *The numbers and percentages of web site interface pages requested in exploratory sessions by expMi and expEn.*

The majority (84.0%) of expMi session requests were requests for the Niupepa home page, including links to different sections on the home page. This preference is to be expected because the home page is usually the first page accessed by clients exploring a web site. The second highest preference for the

expMi sessions, at 8.9%, was the About English Abstracts page, which initially suggests that expMi users are interested in reading about the English Abstracts. However, the term used in the link to this page, ‘he rāpopototanga Reo Pākehā’, could also be translated as ‘An English Summary’. Hence, expMi users could have been mistakenly accessing this link in the hope of viewing a summary of the Niupepa collection in English. By contrast, the second highest preference in expEn sessions, at 6.8%, is the NZDL home page, suggesting that exploratory clients in English were also interested in information about the NZDL.

3.3.4 Summary—Exploratory Sessions

Exploratory sessions in te reo Māori (expMi) were short sessions, mostly consisting of two or three requests, and half of the sessions were completed within 30 seconds. While the median apparent read times of interface pages in expMi sessions was 19 seconds, the average apparent read time was 146 seconds, suggesting that there were many expMi sessions where the clients were taking longer than two minutes to read and explore the Niupepa interface in te reo Māori. The home page, and sections within the home page, was the interface page mostly requested by expMi users. The About English Abstracts explanation page was also significantly requested by expMi sessions; however, this may have been a result of an ambiguous translation of this hyperlink.

Exploratory sessions only request pages from the Niupepa interface; no documents are requested and no queries are submitted. Aside from the details listed in this summary, it is difficult to derive any further information about how te reo Māori was used by exploratory (only) clients of the Niupepa digital library.

This summary leads us into the next part of the analysis—how te reo Māori was used by clients who were actively using the Niupepa digital library to access information. Usage of the Niupepa digital library by clients in te reo Māori (usgMi clients) will be analysed in three parts, interface behaviours, document access behaviours and query behaviours. The first part, an analysis of how usgMi clients used the Niupepa interface, begins in the following section.

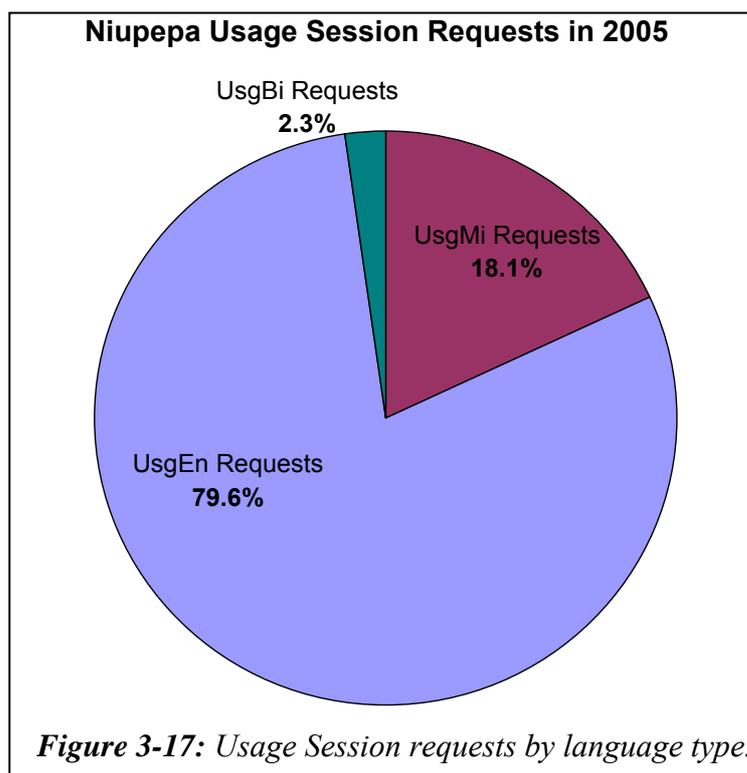
3.4 Niupepa Interface Behaviours

This section analyses how the Niupepa interface was used by reo Māori clients in usage sessions. A usage session is a session where the client actively engages with the Niupepa web site to view documents in the Niupepa collection (for a description of how usage sessions are defined see §2.4.2). Usage session data is the most important data to analyse as it explicitly examines the key target group, clients who actively use the Niupepa digital library.

There were 172,905 usage requests recorded in 6,714 sessions for the Niupepa collection in 2005. Of these requests, 33,540 (19.4%) were when the interface was set to reo Māori and 139,365 (80.6%) were when the interface was set to English. Usage sessions lasted approximately ten times longer and had approximately ten times more requests than the exploratory sessions.

The usage session requests were then subdivided by language type (as explained in §2.4.3). There were 31,287 requests in 1,702 usage sessions when the interface language was mostly set to reo Māori (usgMi). There were 137,697 requests in 4,952 usage sessions when the interface language was mostly set to English (usgEn). There were 3,921 requests in 90 usage sessions when the interface language was defined as bilingual (usgBi). A count of the usgMi requests shows the interface language set to te reo Māori in 98.5% of the requests. A count of the usgEn requests shows the interface language set to English in 98.8% of the requests. A count of the usgBi requests show the interface set to English in 73.5% of the requests and set to reo Māori in the remaining 26.5%.

The usage session requests for each language type were calculated as percentages and are displayed in *Figure 3-17*.



Statistics generated from TLA were used to examine seven interface behaviours exhibited by clients who used the Niupepa digital library in te reo Māori. The interface behaviours examined were:

- the length of usgMi sessions
- the number of return usgMi sessions
- the origin of usgMi sessions
- the interface pages requested in usgMi sessions
- the apparent page read times of interface pages in usgMi sessions
- the frequency of interface language switching in usgMi sessions
- the position of interface language switching in usgMi sessions.

These seven interface behaviours comprise the next seven sub-sections and provide a description of how te reo Māori clients used the Niupepa interface.

3.4.1 How Long Were Usage Sessions?

Two measurements were used to establish the length of usgMi sessions; duration in time (seconds and minutes) and length in number of requests. Statistics

generated from TLA on the Niupepa digital library are listed to show how long usgMi sessions of the Niupepa digital library were in 2005. These statistics are listed alongside length statistics of usgEn sessions and usgBi sessions.

How Long in Time Were Usage Sessions?

UsgMi sessions, like all usage sessions, generally did not last for a long time. The shortest usage sessions lasted less than a second, the longest for almost six hours and the average was 20 minutes. The median time for usage sessions was 6 minutes.

The durations of usgMi sessions, usgEn sessions and usgBi sessions were grouped into the following time intervals; 0–10 seconds, 11–60 seconds, 1–3 minutes, 3–15 minutes, 15–30 minutes, and more than 30 minutes. The numbers of sessions that fitted into these time categories were calculated as percentages for comparison purposes and graphed in *Figure 3-18*.

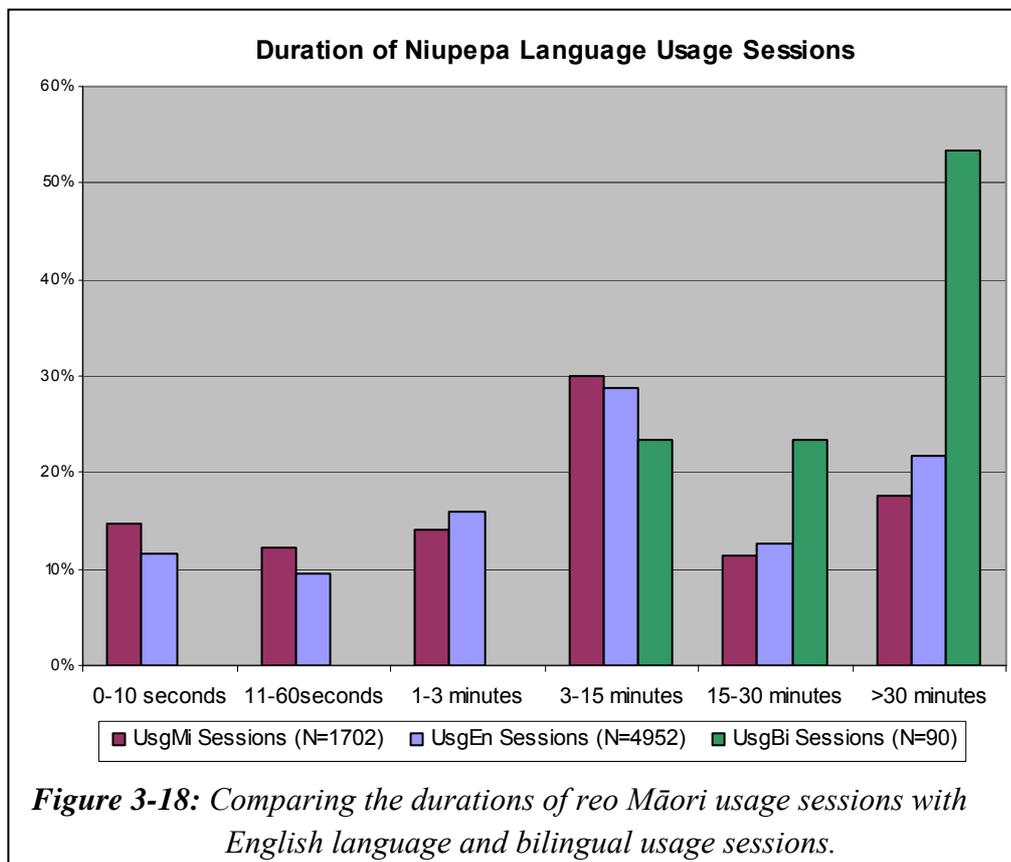


Figure 3-18 shows that usgMi sessions and usgEn sessions appeared to have similar session durations, while the usgBi sessions were all longer sessions. However, when chi square analysis is undertaken on the session lengths of usgMi and usgEn sessions it generates a p value of $3.0 * 10^{-6}$. This indicates that the difference is statistically significant. UsgMi sessions had a higher proportion of shorter sessions (in the 0–10 second and 11–60 second category) while usgEn sessions had a higher proportion of longer sessions (in the 15–30 minute category and >30 minute category). The average usgMi session duration was 16.8 minutes; the average usgEn session duration was 21.3 minutes. Although *Figure 3-18* appears to show that usgMi sessions were of similar duration to usgEn sessions, the figures indicate that usgMi sessions were statistically shorter than usgEn sessions.

The statistics for usgBi sessions indicated that these sessions were almost three times longer in average length than usgMi sessions. One of the criteria for determining a usgBi session was that these sessions must be at least 10 requests long. This criterion alone ensures that usgBi sessions are long. It is also difficult to deduce usgBi characteristics with the same amount of confidence as usgMi and usgEn characteristics because the number of sessions is comparatively low at only 90 sessions.

How Long in Number of Requests Were Usage Sessions?

Usage sessions consisted of requests for pages from the Niupepa interface and at least one request for an actual document from the Niupepa collection. The shortest usage sessions involved just two requests and sessions of just two requests were the most common length. The longest session involved 598 requests. There was an average of 25.6 requests per usage session.

To illustrate the requests recorded in usgMi sessions the numbers of requests per usage session were grouped into the following divisions; 2 requests, 3–4 requests, 5–8 requests, 9–20 requests, 21–40 requests and 41+ requests. The numbers of sessions per request count category were calculated as percentages and are displayed in *Figure 3-19*.

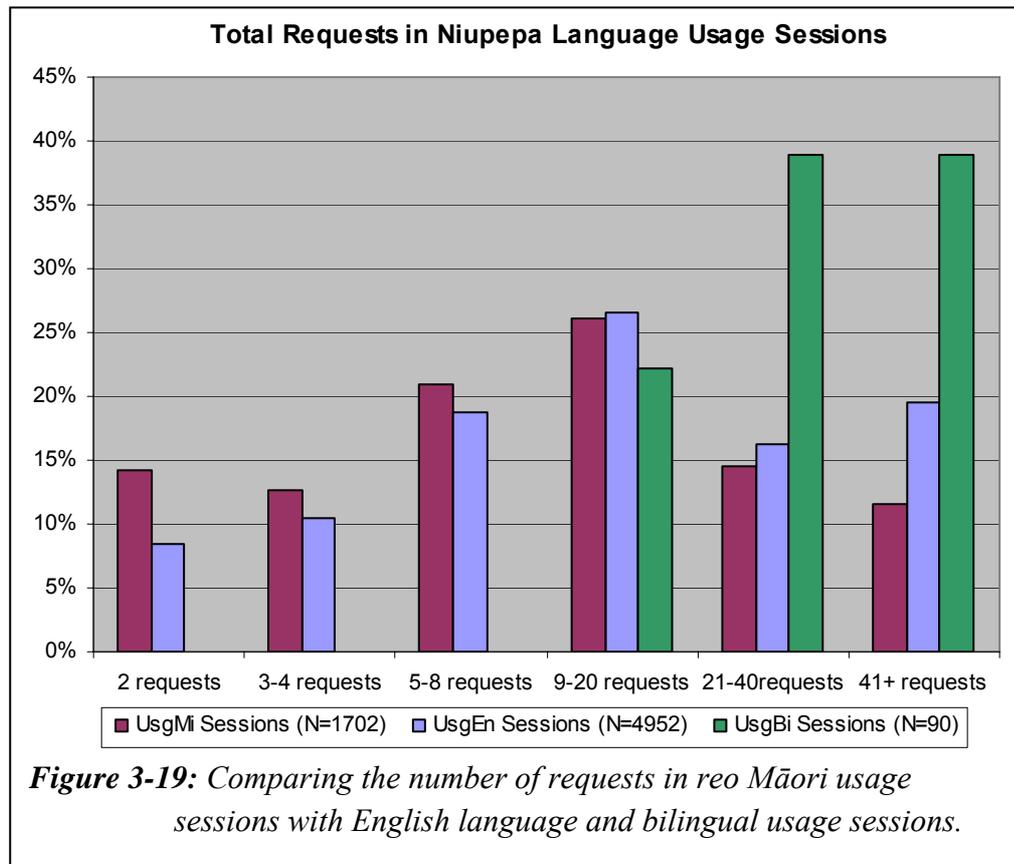


Figure 3-19 shows that usgMi sessions had the highest percentages of short sessions; sessions with only 2 requests, 3–4 requests and 5–8 requests. The usgEn sessions had a higher proportion of longer sessions than the usgMi sessions but the usgBi sessions had the highest proportion of longer sessions.

UsGMi sessions had an average of 18.4 requests per session and a median of 9 requests per session. The usgEn sessions had an average of 27.8 requests per session and a median of 12 requests per session. Thus the usgMi sessions, on average, consisted of a third fewer requests than the usgEn sessions.

The statistics for usgBi sessions indicated that these sessions had almost twice as many requests as usgMi and usgEn sessions. Again, this high number of requests may simply be because of the criteria for determining a usgBi session, that these sessions must be at least 10 requests long.

3.4.2 How Long to Read Interface Pages?

Reading of interface pages has been defined as apparent page reading (see §2.5.1). The apparent page read time is the length of time from when one interface page of the web site is requested until another interface page or document is requested. The apparent read times in usgMi, usgEn, usgBi sessions and in all usage sessions were calculated and descriptive statistics have been displayed in *Table 3-6*.

UsgMi clients, on average, took 51.5 seconds to read the interface pages in usage sessions. This apparent page read time is almost 1/3 faster than it took expMi clients to read interface pages (145.6 seconds) in exploratory sessions (*Table 3-4*).

	usgMi	usgEn	usgBi	Total
Count	1339	5810	542	7691
Average (seconds)	51.5	48.7	71.8	50.8
Standard Deviation	162.1	187.2	223.0	186.0
Mode (seconds)	4	4	4	4
Shortest (seconds)	0	0	0	0
Lower Quartile (seconds)	5	6	6	6
Median (seconds)	10	12	15	12
Upper Quartile (seconds)	31	32	40	33
Longest (seconds)	2131	3237	2367	3237

Table 3-6: Summary statistics comparison of usgMi, usgEn, usgBi and total apparent page read times of interface pages in usage sessions.

The average apparent usgEn page read time (48.7 seconds) seems similar to the average usgMi apparent read time (51.5 seconds). However a chi square analysis comparing the usgMi and usgEn apparent read times gives a p value of $4.8 * 10^{-11}$, indicating that the usgMi apparent page read times are significantly longer.

3.4.3 How Many Clients Returned?

TLA of the Niupepa web log can give indications of how often clients returned to the Niupepa digital library. The results generated assume that clients always have cookies enabled, the cookies are not routinely deleted, clients always log on using the same account on the same computer, and that there is minimal to nil network and local browser caching involved. These assumptions are not valid for some users in some sessions. However, analysing return statistics can still give an indication of how often usgMi clients returned to use the Niupepa digital library.

Statistics from TLA indicated that from a total of 6,117 identifiable clients there were at least 283 (4.6%) clients who undertook return sessions. These 283 clients made 33,099 (19.1%) requests for pages and documents from the Niupepa web site. These figures indicate that only a small percentage of clients were identified as return clients but these clients generated a high proportion of the requests.

Of the usgMi clients 82 (5.2%) undertook return sessions requesting 4,316 (13.8%) pages and documents. There were also 192 (4.3%) usgEn clients who undertook return sessions requesting 27,112 (19.7%) pages and documents. There were 52.6 requests submitted per returning usgMi client compared with 141.2 requests submitted per returning usgEn client, indicating that the return usgEn clients submitted more requests than the return usgMi users.

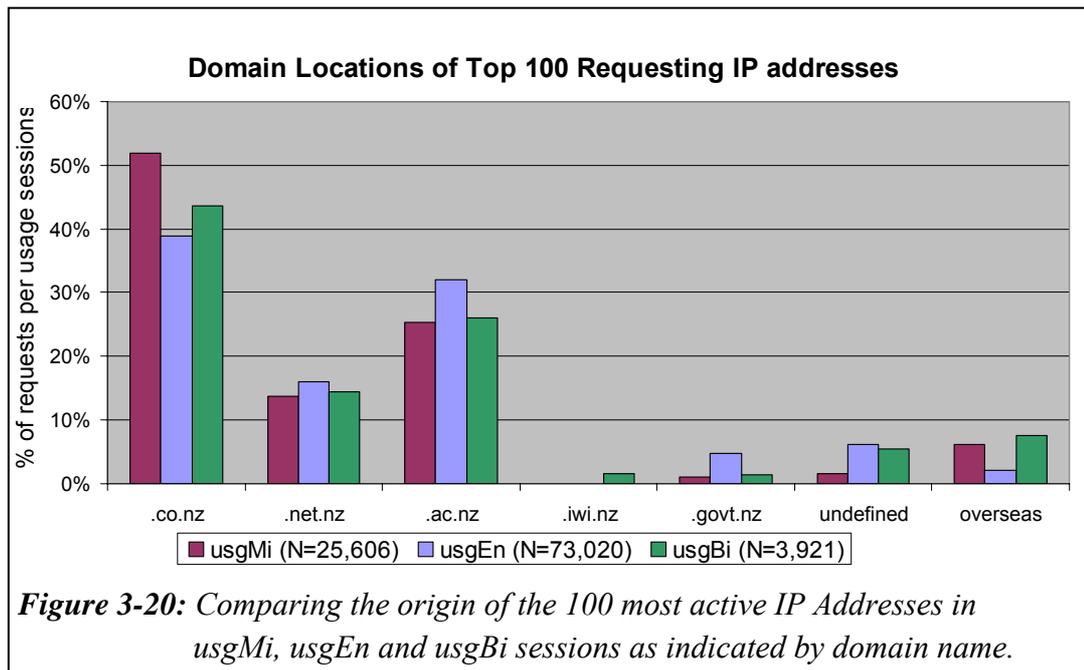
None of the clients classified as usgBi clients undertook return sessions as a usgBi client but 13 (14.4%) of these clients were recorded as other user types. There were two usgBi clients who were also recorded as usgMi clients and 11 usgBi clients who were also recorded as usgEn clients. These figures suggest two usage behaviours. First, switching the language of the interface is an activity undertaken only once by users. Second, a proportion of usgEn clients are prepared to try the interface in the medium of te reo Māori.

3.4.4 Where Did Usage Sessions Originate?

Information about the origin of requests can be derived from TLA. The second argument recorded in the web logs with each request was the unique IP Address (Internet Protocol address) of the computer or network device that submitted the request. The IP Addresses can be examined by a process called reverse DNS lookup, which returns information about the computer/network device including domain name and location. Counts can then be undertaken to determine the number of requests submitted per domain name and per location.

The IP Addresses of the requests recorded in usgMi sessions were examined using reverse DNS lookup to see if any trends in domain name could be determined. Eight distinctive domain name types were identified, `.co.nz`, `.net.nz`, `.ac.nz`, `.iwi.nz`, `.govt.nz`, *undefined* and *overseas IP*. The first six have a common first level domain name (i.e. `.nz`) but are different in the 2nd level domain name; the seventh domain type, *undefined*, represents those IP Addresses which did not return a valid domain name when queried; and the eighth domain type is a grouping of all the first level domain names which did not equal `.nz` and were subsequently grouped as being requests from overseas.

In total 3,183 IP Addresses submitted requests to the Niupepa digital library in 2005. Determining individual domain names on all 3,183 IP Addresses is beyond the scope of this thesis. Consequently, only the 100 IP Addresses submitting the most requests in usgMi sessions were examined. To provide a comparison the 100 IP Addresses submitting the most requests in usgEn sessions and the total 69 IP Addresses making requests in the usgBi sessions were also examined. The percentages of requests in these usage sessions are displayed in *Figure 3-20*.



The usgMi requests had a similar origin to the usgEn and usgBi requests; most requests came from the .co.nz domain, followed by the academic community the .ac.nz domain, with the third largest domain being the .net.nz group. These results suggest that the use of te reo Māori in the Niupepa digital library was not limited to one particular domain type.

The .co.nz group has been defined as organisations pursuing commercial aims and purposes (Office of the Domain Name Commissioner, 2005) and the .net.nz group has been defined as organisations and service providers directly related to the New Zealand (Aotearoa) Internet. The two domain names have been used for similar purposes by Internet Service Providers in Aotearoa and if these two names are grouped together they account for an average of 59.6% of the IP Addresses making requests in each usage session group.

The .ac.nz domain name is the origin of the second largest group of IP Addresses and is analysed in more depth in the following sub-section. The IP Addresses which indicate requests from overseas also reveal some interesting information, as discussed in a following sub-section.

There was one IP Address that originated from the second level domain name of `.iwi.nz`. This domain name is reserved for traditional Māori tribes. It seems unusual that this domain name only appears in the usgBi sessions. Perhaps if more than the top 100 IP Addresses were examined in the usgMi and usgEn sessions then the `.iwi.nz` second level domain name may have appeared in those sessions as well.

If we examine the second level domain name of `.govt.nz` we see that there are three times as many IP Addresses making request in English usage sessions as there are in te reo Māori or bilingual usage sessions. This is, perhaps, an indication of a low number of reo Māori speakers working in the New Zealand Government.

Originating Academic Institutions

The requests recorded from the academic second level domain name were isolated and grouped by the originating academic institution. These academic institutions are displayed in *Figure 3-21* sorted by location and listed from north to south.

Nearly half of all Māori live in Auckland, Waikato and the Bay of Plenty (Te Puni Kōkiri, 2001b). As the Niupepa collection is primarily a reo Māori resource we would expect most of the requests to come from this area and this expectation is indeed reflected in *Figure 3-21*.

Statistics provided by the Ministry of Education for 2005 (Alosius, 2007) show that 3,613 Māori students enrolled at Massey University, 3,028 Māori students at Waikato Institute of Technology, 2,802 Māori students at the University of Auckland and 2,239 Māori students at the University of Waikato. These figures represent the four institutions displayed in *Figure 3-21* that have the highest number of Māori students enrolled. The Eastern Institute of Technology had the highest percentage of Māori student enrolments at 22.5%, followed by Waikato University at 16.7%, the Southern Institute of Technology at 14.9% and the

Waikato Institute of Technology at 14.7%. All other education providers had less than 10% of Māori student enrolments.

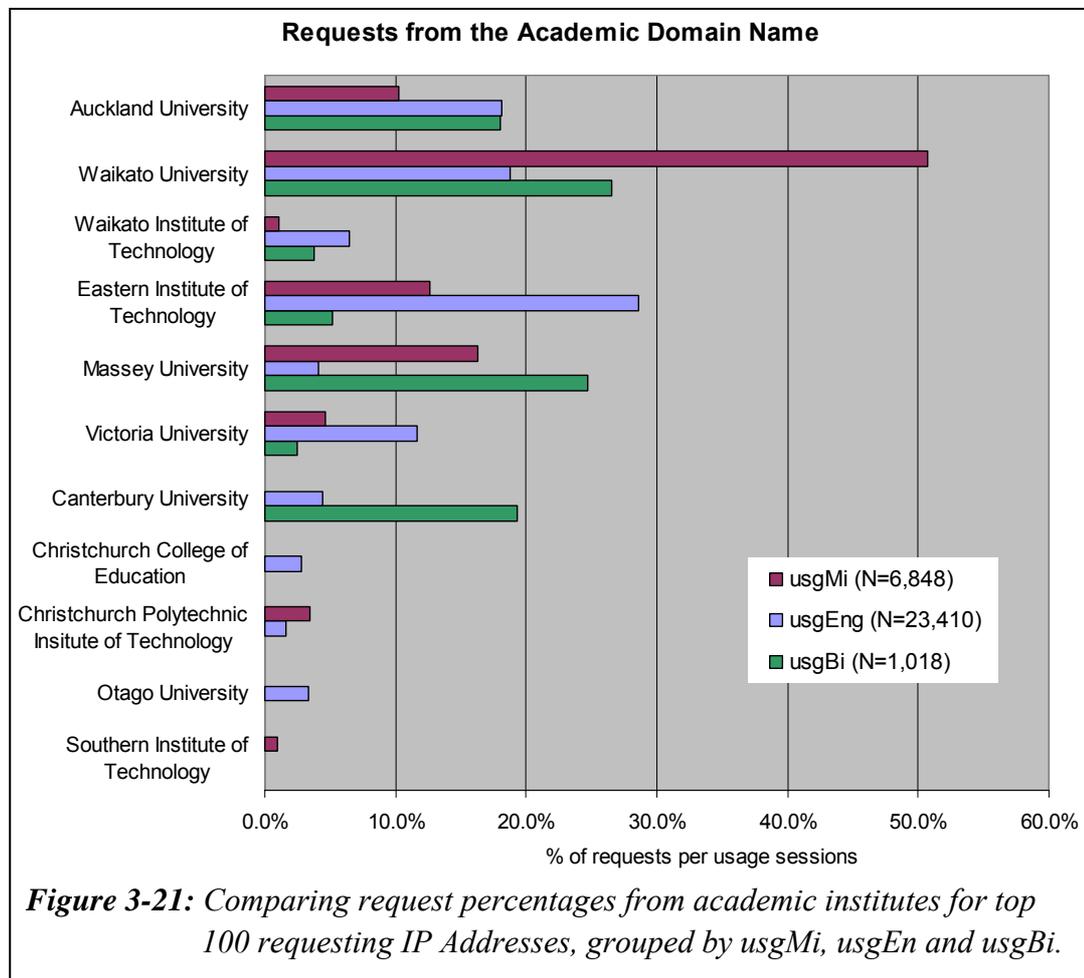


Figure 3-21 indicates more requests coming from the northern tertiary institutions, Auckland University down to Massey University. This trend mirrors the statistics provided by the Ministry of Education which state that two thirds (65.5%) of the Māori student population is studying at these institutions.

There was a proportionally higher number of requests originating from the University of Waikato. As Waikato University is the host institute of the Niupepa digital library perhaps the digital library is more widely known and, as a result, more widely used at this university.

Originating Countries

The requests from 1st level domain names that did not equal .nz were examined to determine which countries were making the most requests to the Niupepa collection. Overseas IP Addresses that made 20 or more requests to the Niupepa collection were collated and then plotted in *Figure 3-22*. It should be noted that requests from Hawai‘i were separated from the US requests even though Hawai‘i is a state of the United States.

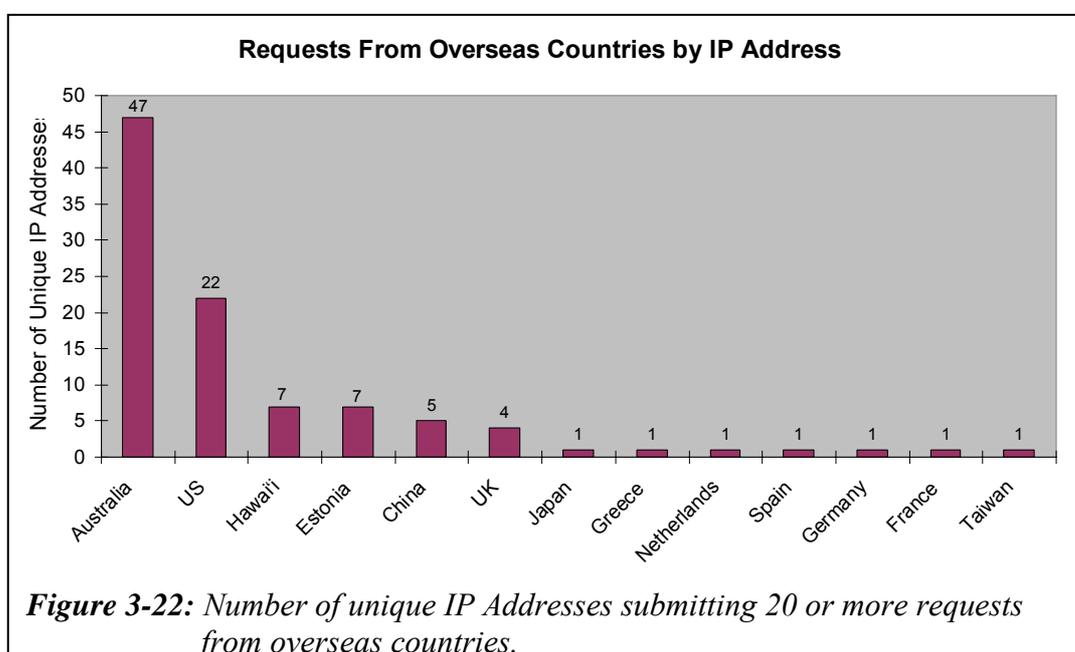


Figure 3-22 shows some interesting characteristics. First, it highlights that the most overseas usage requests in 2005 were from Australia, with 47 separate IP Addresses (submitting 2,697 usage requests). Given that estimates as early as 2001 suggested over 90,000 Māori (over 15% of all Māori) live in Australia (Bedford et al, 2004) this high interest in the Niupepa web site by people in Australia, presumably Māori living there, is not surprising. Following Australia, the next country using the collection was the US, with 22 unique IP Addresses submitting 847 usage requests. Next was Hawai‘i, a state of the US, with 7 separate IP Addresses submitting 562 usage requests. A significant number of requests from Hawai‘i is not surprising given that the Hawaiian and Māori people

are both Pacific Island nations sharing a similar language and culture, and that both nations have been involved with building digital libraries based on legacy newspapers.

Perhaps what is surprising though is the reasonably high appearance in the graph of Estonia, with 7 unique IP Addresses making 441 requests in usage sessions. As 403 of those requests occurred in April and May, one can speculate that these Estonian usage sessions represent a particular concentrated effort from a person, or group of people working together, to research some aspect of Aotearoa's encounter history, some aspect of Aotearoa's indigenous language or history or, perhaps, some aspect of using the Greenstone software to build digital libraries.

Most (82.0%) of the usage requests displayed by IP Addresses in *Figure 3-22* are requests from usgEn sessions. There were also 14.8% recorded from usgMi sessions and the remaining 3.2% were from usgBi sessions. Countries with 75% or more of requests in English language sessions included Australia, US, Estonia, China, UK, Japan, Spain, Germany and Taiwan. Countries that appeared prominent in the reo Māori usage sessions were Hawai'i, with 351 requests representing 62.5% of its usage requests and Greece, with 31 requests at 64.6% of its usage requests. In addition, the 21 requests from the solitary French IP Address were all recorded as reo Māori usage session requests.

Only one country, the Netherlands, appeared prominent in bilingual usage session requests with 30 (65.2%) of its 46 requests recorded as bilingual usage sessions and the other 16 being recorded as English usage sessions. These figures suggest that clients from Hawai'i, the client from Greece, and the client from France were willing to undertake activities on the Niupepa digital library in te reo Māori. The figures also suggest that the client from the Netherlands switched languages often when using the web site.

3.4.5 Which Interface Pages Were Requested?

The interface pages requested by clients who use te reo Māori are examined in this sub-section. Note: a page refers to an interface page of the Niupepa web site rather than a newspaper document within the Niupepa collection.

The numbers and percentages of interface pages requested in usgMi sessions are listed in *Table 3-7*. Also listed for comparative purposes are the numbers and percentages of pages requested in usgEn sessions and usgBi sessions.

	usgMi		usgEn		usgBi	
Total pages	1390		6080		551	
Niupepa home page	1127	81.1%	4931	81.1%	413	75.0%
NZDL home page	18	1.3%	121	2.0%	9	1.6%
About English Abstracts	83	6.0%	349	5.7%	26	4.7%
help page	48	3.5%	150	2.5%	25	4.5%
preferences page	114	8.2%	529	8.7%	78	14.2%
about greenstone page	0	0.0%	0	0.0%	0	0.0%
unknown page	0	0.0%	0	0.0%	0	0.0%

Table 3-7: The numbers and percentages of interface pages requested by usgMi, usgEn and usgBi clients.

Table 3-7 shows the majority of usgMi requests, at 81.1%, were requests for the Niupepa home page which, as a home page, can be expected. While the second most requested interface page on the expMi (exploratory) sessions was the About English Abstracts page, here in usage sessions the second most requested page was the Preferences page at 114 (8.2%) requests.

This characteristic is also expected when clients are actively using the collection as they are likely to request the Preferences page to alter language, format and searching preferences. However, while this is a significant proportion of the interface pages requested in usgMi sessions, it is not a significant proportion of

the total requests. There were 31,287 total requests in usgMi sessions; 114 requests for the Preferences page represent less than 0.4% of these total requests.

The percentages of interface pages accessed by reo Māori and English sessions were similar. When chi square analysis is undertaken on the counts of usgMi and usgEn interface pages a p value of 0.11 is generated. This value indicates that there is no significant difference in the type of interface pages that are being requested in usgMi and usgEn sessions. Thus we can conclude that clients who used the Niupepa digital library in te reo Māori were requesting the same proportion of interface pages as those who used the Niupepa digital library in English.

3.4.6 How Often Was the Interface Language Switched?

In general, usgMi clients showed a reluctance to switch the language of the interface. Of 1702 usgMi sessions only 142 (8.3%) switched the interface language; most of these switches, 114 of 193 (59.1%), were a switch from English to reo Māori.

By comparison, 1067 (21.6%) of 4952 usgEn clients switched the interface language; most of these switches, 1058 of 1244 (85.1%), were a switch from reo Māori to English. By definition a usgBi session contains 3 or more language switches; there were 90 usgBi sessions involving 352 interface language switches, 154 (43.7%) of these were a switch from English to reo Māori and 198 (56.3%) a switch from reo Māori to English.

3.4.7 When Was the Interface Language Switched?

The usgMi sessions were divided into 10 equal portions and the portion in which the language switch occurred in each session was noted. These positions are plotted in *Figures 3-23* and *3-24*. The positions that the language switch occurred in usgEn and usgBi sessions are also plotted in *Figures 3-23* and *3-24*.

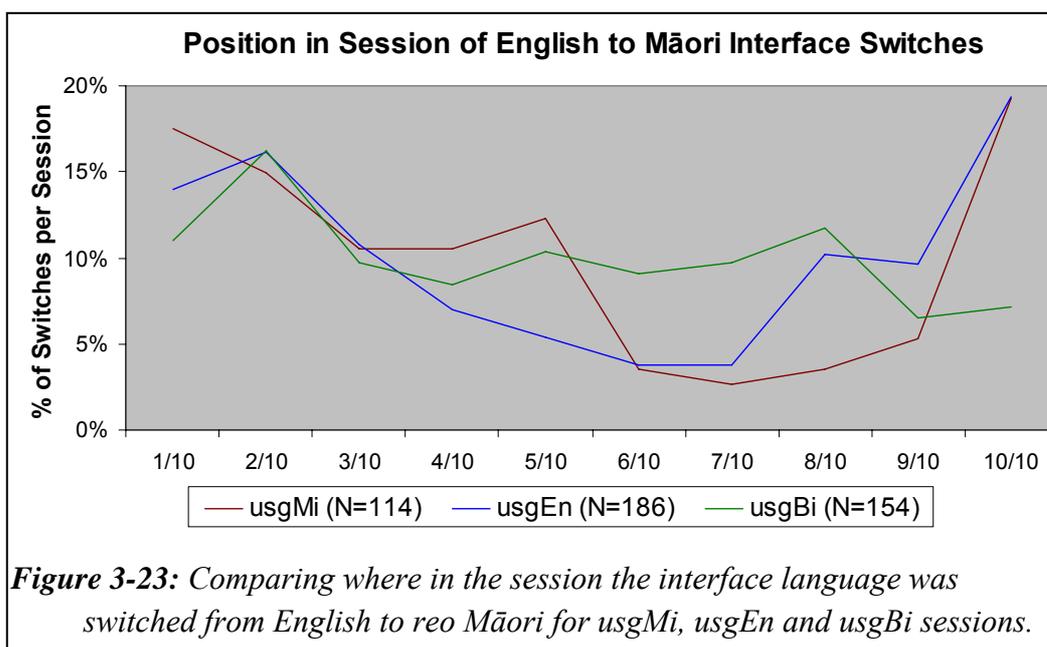
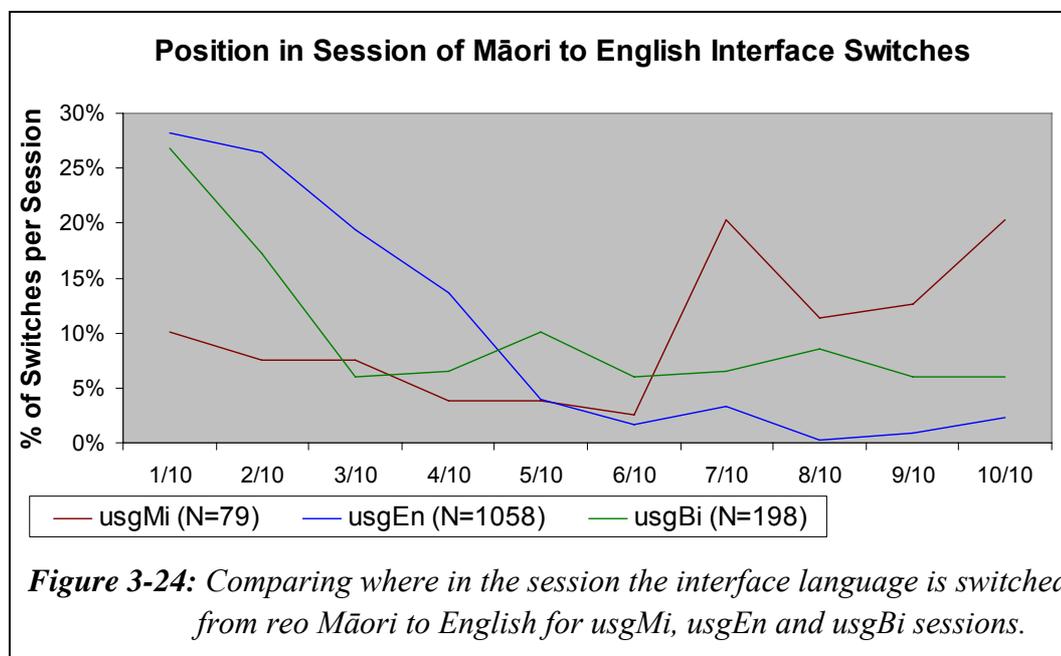


Figure 3-23 shows that clients in usgMi sessions mostly switched the interface language from English to reo Māori in the first half of the session or the very last tenth of the session. There were only 114 interface language switches to reo Māori in usgMi sessions. This low number of switches to reo Māori by usgMi clients suggests that many of the 1702 usgMi sessions were occurring because either the default language of the web site was set to reo Māori or the clients were accessing the web interface through a bookmark that included the default interface language setting of reo Māori.

The usgEn sessions had a higher number (186) of interface switches into reo Māori than the usgMi sessions (114). This number indicates that many usgEn clients must be confident enough in their reo Māori abilities to try the interface in te reo Māori. As most of these switches occurred in the first 3/10ths of the

sessions it perhaps indicates usgEn clients exploring the reo Māori version of the interface.



In contrast, *Figure 3-24* shows the switch of the interface language from reo Māori to English by usgMi clients occurs mostly in the final 4/10ths of usgMi sessions. The position of this switch could suggest a frustration in the ability to use the interface in te reo Māori, although the number of switches is low, 79, representing less than 5% of the total 1702 usgMi sessions.

The largest number of interface switches was from reo Māori to English in usgEn sessions. There were 1058 switches to English in the 4952 usgEn sessions. These figures indicate that more than 1 in 5 usgEn clients switched the interface language to English. As these switches were mostly undertaken at the start of a session it suggests that setting the default language of the Niupepa interface to English would be welcomed by clients who use the Niupepa digital library in English. The default language setting of the web site is comprehensively discussed in Chapter 5.

3.4.8 Summary—Interface Behaviours

This section analysed how the Niupepa interface was used by te reo Māori clients in usgMi sessions. Statistics drawn from TLA examined seven aspects of using the Niupepa interface in te reo Māori.

The statistics reveal that on average a usgMi session lasted for 30 minutes and consisted of 18 requests. The average apparent read time of an interface page by usgMi clients was 52 seconds. Only 6% of the usgMi sessions were identified as return sessions; however, these return sessions were responsible for 14% of the total usgMi requests. More than half of the usgMi requests from an academic institution in Aotearoa originated from Waikato University. There were twice as many usage requests from Australia as there were from any other overseas country. The home page was the most popular interface page requested by te reo Māori clients, at 81% of all interface page requests, followed by the Preferences page at 8%. The language of the interface was only switched 142 times in 1702 usgMi sessions. When the language of the interface was switched, it was switched more often from English to reo Māori (59%) and mostly in the first half of the session.

This brief summary completes the examination of how the Niupepa interface was used by clients in te reo Māori. Usage analysis of the Niupepa digital library by usgMi clients was undertaken in three parts; an analysis of how the interface was used, of how documents were accessed and of how queries were undertaken. Having completed this analysis the next stage is to examine how documents were accessed by usgMi clients. This analysis begins in the next section.

3.5 Document Access Behaviours

This section analyses how te reo Māori clients accessed Niupepa documents in usage sessions. By definition (see §2.4.2) if a client requested a document in a session then that session was regarded as a usage session. A document request is defined as a request for some document within the Niupepa digital library that represents real content, rather than something that provides access to content, such as an interface page. Typically this resource is a newspaper page but it could also be a request for an English Abstract (for an example see *Figure 3-9*) or a Bibliographic Commentary (for an example see *Figure 3-10*).

TLA was used to provide statistics on seven document access behaviours exhibited by clients who used the Niupepa digital library in te reo Māori. The document access behaviours examined were:

- the methods used to access documents in usgMi sessions
- the document formats favoured in usgMi sessions
- the document language favoured in usgMi sessions
- the publication years most requested in usgMi sessions
- the propensity for documents with associated text files
- the propensity for documents with an associated English Abstract
- the apparent read times of documents by usgMi clients.

These seven document access behaviours comprise the next seven sub-sections and provide a thorough description of how te reo Māori clients accessed documents of the Niupepa digital library.

3.5.1 How Were Documents Accessed?

There were three methods available to access documents of the Niupepa digital library; selection of a document following a query activity, selection of a document from a list of publication series, and selection of a document from a list

based on dates. When a client requested a document by one of these methods, say a query, and then proceeded to browse documents in the collection, the subsequently browsed documents were all recorded as being selected from the query activity.

Results from TLA indicated the methods used by usgMi clients to access Niupepa documents. The counts and percentages are displayed in *Table 3-8*, along with counts and percentages of how documents were accessed by usgEn and usgBi clients.

How Documents Were Accessed								
	usgMi		usgEn		usgBi		Total	
Total	22,248	100%	102,040	100%	2,272	100%	151,260	100%
From query	17,282	77.7%	83,221	81.6%	1,595	70.2%	102,098	80.7%
From series	2,878	12.9%	6,754	6.6%	233	10.3%	9,865	7.8%
From date	1,641	7.4%	9,860	9.7%	346	15.2%	11,847	9.4%
Uncertain	447	2.0%	2,205	2.2%	98	4.3%	2,750	2.2%

Table 3-8: Counts and percentages comparing how documents were accessed by usgMi, usgEn and usgBi clients and the total usage session clients.

Table 3-8 shows that the query activity (see *Figure 3-3*) was the most popular method used by usgMi clients to access Niupepa documents. The usgMi clients showed a second highest preference for accessing documents from the series list (see *Figure 3-4*) and the least popular method was accessing the document from a list based on dates (see *Figure 3-5*).

Clients in all three usage sessions preferred to access documents from a query activity. However, the usgEn and usgBi clients differed from the usgMi clients in that the second preferred method for accessing documents was from a list based on dates, while the least preferred method was from the series list.

3.5.2 Which Document Formats Were Favoured?

The newspaper pages stored in the Niupepa digital library can be presented to the client in one of three formats; extracted text generated from the OCR process, a preview image and a full size image. Examples of these formats are shown in *Figures 3-6, 3-7, and 3-8*. *Table 3-9* lists the numbers and percentages of these three formats, along with a fourth undefined listing where the format of the page was not recorded correctly in the web log. The undefined requests are most likely caused by bookmarked pages being requested following software changes to the Niupepa web site.

Formats of Requested Documents						
	usgMi		usgEn		usgBi	
Total Documents	22248	100%	102040	100%	2272	100%
Extracted Text	13730	61.7%	56042	54.9%	1238	54.5%
Full Size Image	7389	33.2%	39052	38.3%	816	35.9%
Preview Image	999	4.5%	6379	6.3%	211	9.3%
Undefined	130	0.6%	567	0.6%	7	0.3%
Chi square p value	4.3 * 10 ⁻⁸⁰					
Chi square p value	6.8 * 10 ⁻⁸⁷					

Table 3-9: Document formats requested by usgMi, usgEn and usgBi clients.

The document format preferred by usgMi clients was the extracted text format, at 61.7% of all requested newspaper documents. The next preferred format was the full size image at 33.2% and the preview image (4.5%) was the least requested format.

Table 3-9 also shows that all three usage types prefer the Extracted Text format of the document. While the percentages do appear reasonably similar, chi square analysis shows that there is a significant statistical difference in usgMi and usgEn document formats (p value = 4.3 * 10⁻⁸⁰) and also a significant statistical difference in usgMi, usgEn and usgBi document formats (p value = 6.8 * 10⁻⁸⁷).

These p values indicate the usgMi sessions have a higher preference that is statistically significant to request the extracted text formats, while the usgEn and usgBi have a higher statistically significant preference to request the preview and full size image formats.

3.5.3 Which Document Types Were Favoured?

The documents that were available to Niupepa clients consisted of:

- 19,106 newspaper pages written in te reo Māori
- 15,696 newspaper pages written in te reo Māori and English
- 680 newspaper pages written in English
- 313 English Abstract documents that summarise 10 of the newspaper periodicals
- 22 Bibliographic Commentaries written in English with te reo Māori quotations.

To determine which types of documents were most popular to new and general clients, and to remove the influence of heavy users, the top 100 documents requested by unique users were counted. These documents were requested a total of 5,681 times; 1,186 by usgMi clients, 4,219 by usgEn clients and 276 by usgBi clients. The usage session percentages for each type of document count are displayed in *Figure 3-25*.

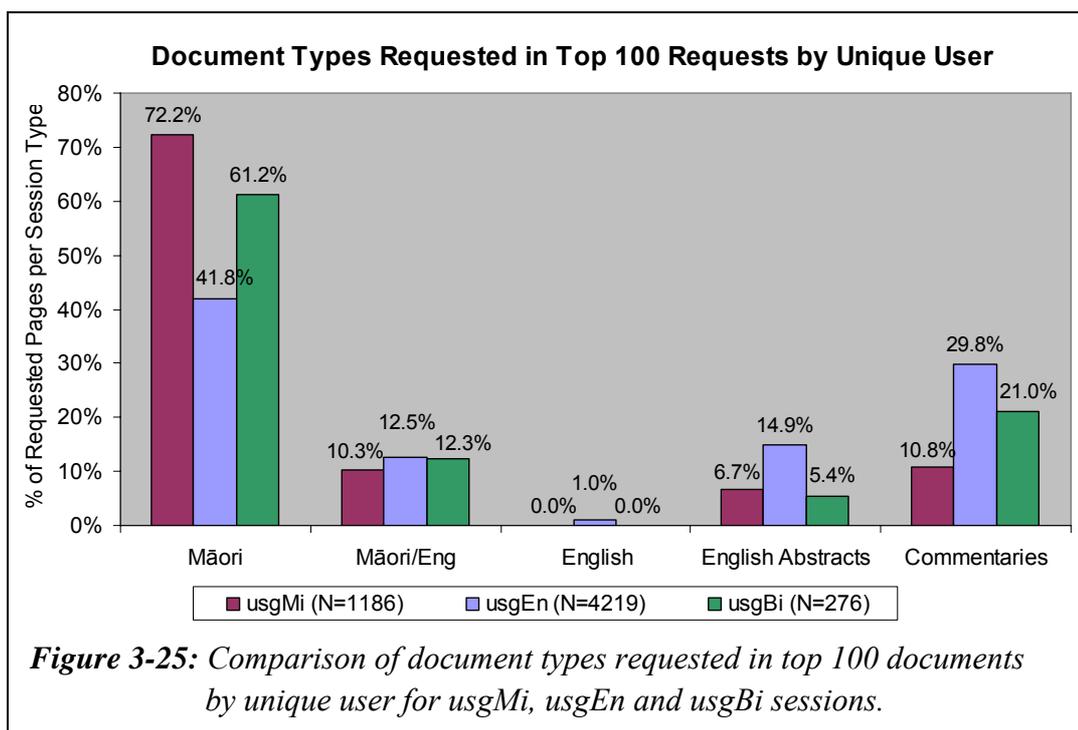


Figure 3-25 shows usgMi clients have the highest percentage of requesting newspaper documents in te reo Māori; 72.2% of the top 100 documents requested by usgMi clients were Niupepa pages written in te reo Māori, compared with 61.2% from usgBi clients and a low 41.8% from usgEn clients.

The usgEn clients had the stronger preference to access documents written in English; the commentary information (29.8%), the English Abstracts (14.9%) and the newspaper pages written in English (1.0%). UsgBi clients requested documents that were between the extremes of those requested by usgMi clients and usgEn clients.

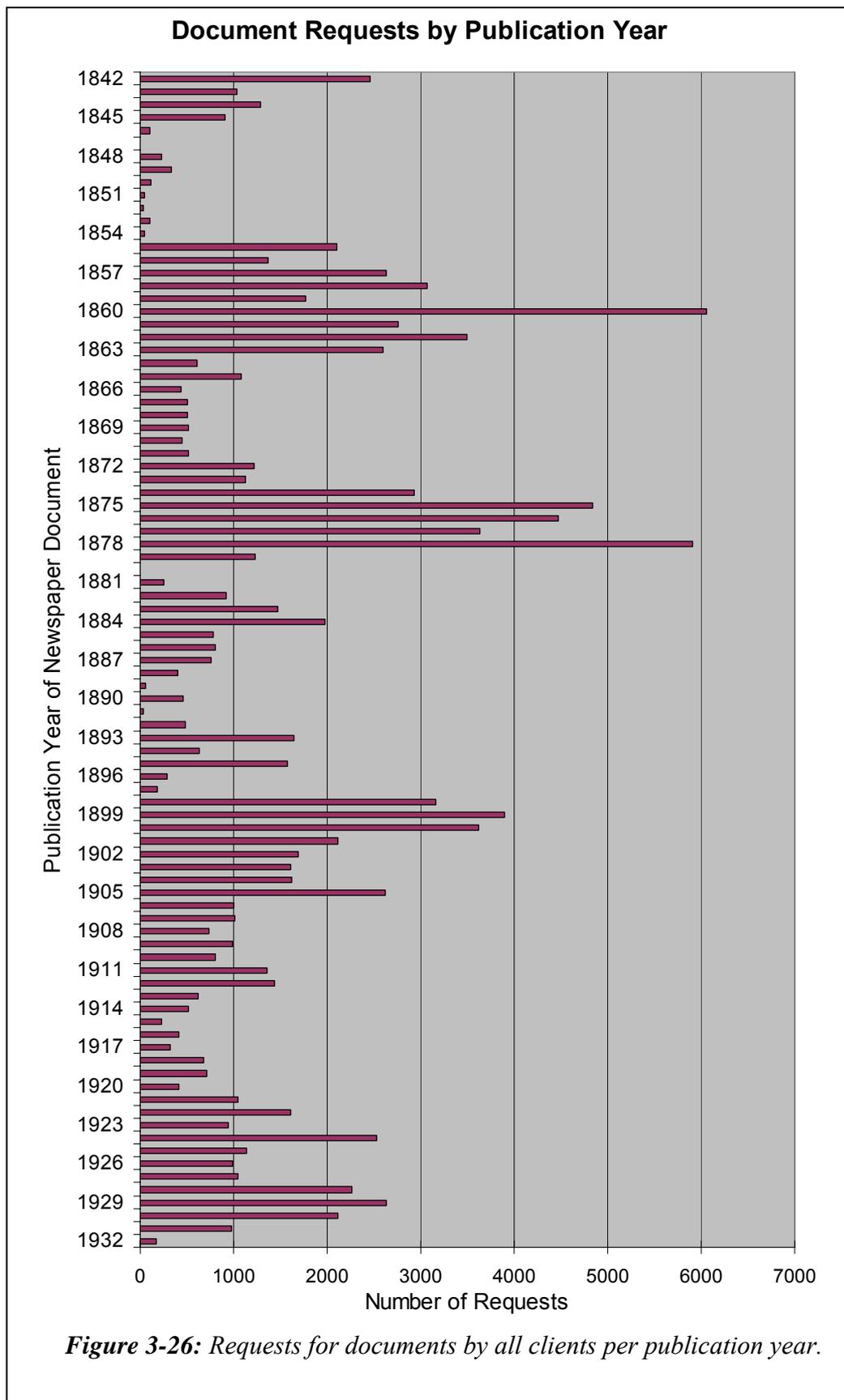
3.5.4 Which Publication Years Were Favoured?

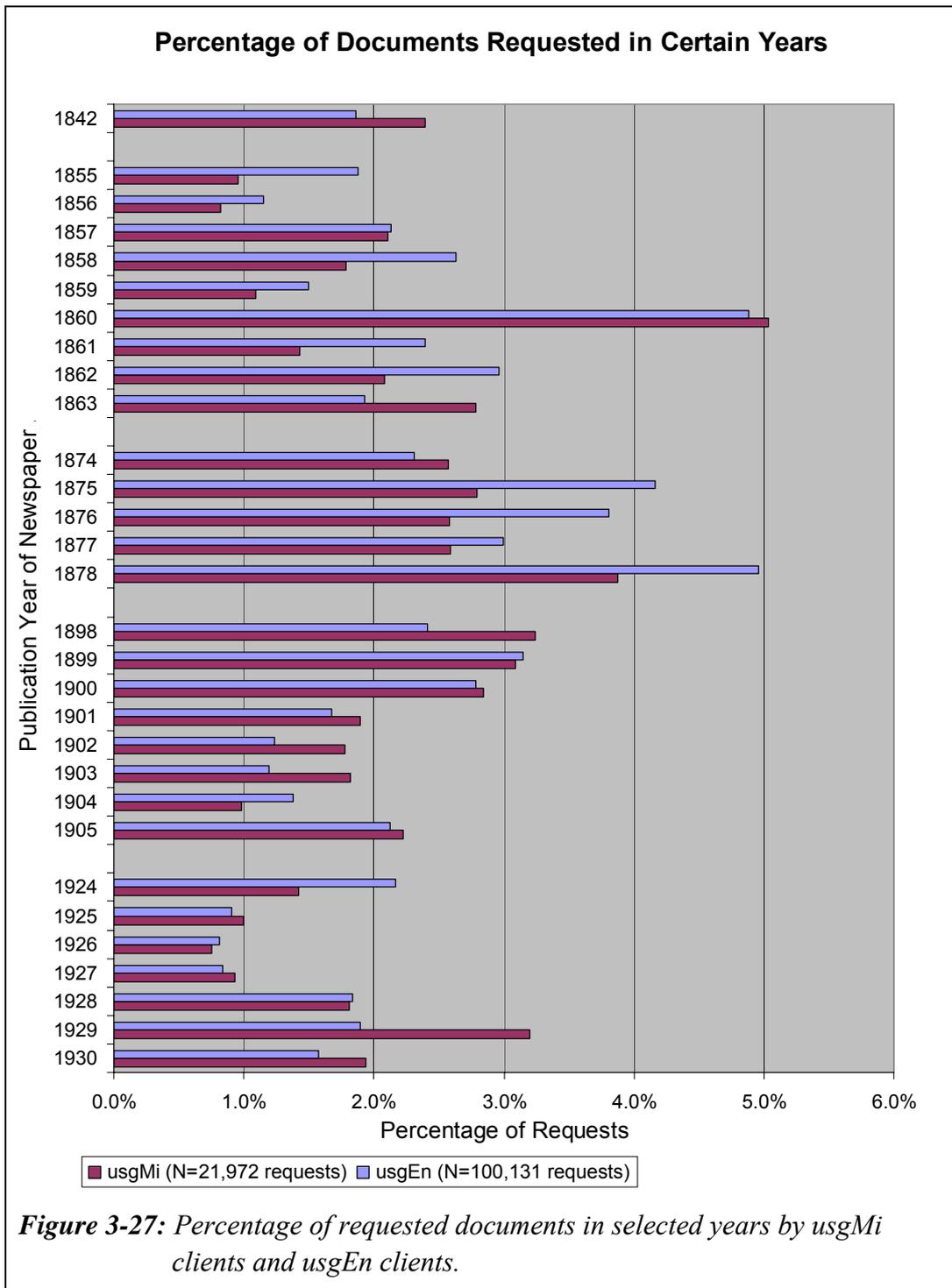
The periodicals in the Niupepa collection were published over a 90 year period, from 1842 to 1932. The count of pages requested from the Niupepa collection per year is displayed in Figure 3-26. This figure shows that publications in certain

years were requested more often by Niupepa clients. The most favoured years were 1842, 1855–1863, 1874–1878, 1898–1905 and 1924–1930.

The popularity of some of these years can be explained. It is not unexpected to see the first year of publications (1842) popular as this is the time period researchers and scholars access to see the first examples of early writing in te reo Māori. There are also English Abstracts available for these early years, making the content understandable to clients with limited or no reo Māori knowledge. The next two time periods, 1855–1863 and 1874–1878 are time periods where bilingual newspapers were published and also English Abstracts are available for those newspapers published only in te reo Māori. Further research is needed to determine why the 1898–1905 and 1924–1930 time periods are so popular.

The most popular dates for which documents are requested have been listed again but this time, rather than the count of the requests, the percentages of the requests are listed. This listing is to determine which dates are important to usgMi clients and also which dates are comparatively important to usgEn clients. Because of the low count of usgBi requests, document dates requested by usgBi clients were not graphed. The usgMi and usgEn request percentages are displayed in *Figure 3-27* which shows that the publication years identified in *Figure 3-26* as important years for clients of the Niupepa collection are also important years for clients who use the Niupepa Collections in te reo Māori. While some small differences are noted in the years usgMi and usgEn clients request documents (for example 1875 and 1921) these differences are not statistically significant.





3.5.5 Documents With Associated Extracted Text Files

If Niupepa documents had associated text files were they requested more often? One would assume this to be true as not only would the documents be

discoverable by browsing series list and by browsing dates, they would also be discoverable by submitting a query. In the Niupepa digital library 93.1% of the documents had an associated extracted text file. Overall there were 118,557 requests (98.6%) for documents with associated text files and 1660 requests (1.4%) for documents without associated text files. On average there were 3.5 requests for every document with a text file and 0.9 requests for every document without a text file. Thus, Niupepa documents with an associated text file were requested 3.8 times more often than Niupepa documents without an associated extracted text file.

This was true for all clients of the Niupepa digital library no matter what language the digital library was used in. A similar analysis was undertaken specifically for usgMi clients and usgEn clients and the results are displayed in *Table 3-10*. The analysis was not undertaken for usgBi sessions as the request count was too low to give meaningful results.

Requests for Documents with Extracted Text Files							
Total Documents		UsgMi Requests			UsgEn Requests		
	Count	Count	%	Req/doc	Count	%	Req/doc
With Text	33,674	20,877	97.6%	0.62	95,609	98.8%	2.8
Without Text	1808	517	2.4%	0.29	1116	1.2%	0.6

Table 3-10: Comparing the number of requests for documents with and without associated extracted text files in usgMi and usgEn usage sessions.

As shown in *Table 3-10* there were 20,877 usgMi requests (97.6%) for documents with text files and 517 usgMi requests (2.4%) for documents without text files. On average there were 0.62 usgMi requests for every document with a text file and 0.29 usgMi requests for every document without a text file, implying that usgMi clients are 2.2 times more likely to request a document with an associated extracted text file.

There were 95,609 usgEn requests (98.8%) for documents with text files and 1116 usgEn requests (1.2%) for documents without text files. On average there

were 2.8 usgEn requests for every document with a text file and 0.6 usgEn requests for every document without a text file, implying that usgEn clients are 4.6 times more likely to request a document with an associated extracted text file.

It is expected that documents with associated extracted text files would be requested more often than documents without. There are more ways in which to request these documents and more forms of the document to request. It would be expected, however, that both usgMi and usgEn would request these documents at an equally high rate. This was not the case. The figures suggest that usgMi clients were less than half as likely to prefer a document with an associated extracted text file.

3.5.6 Documents With Associated English Abstracts

Analysis was undertaken to determine if newspaper pages that had associated English Abstracts were requested proportionally more often than newspaper pages without. One would assume this statement to be true as not only would the documents be discoverable by browsing and searching they would also be discoverable by requesting an English Abstract and following the links from the abstract to the document. There were 22.7% of the Niupepa documents with an associated English Abstract.

There were 0.60 usgMi requests for every document with an English Abstract and also 0.60 usgMi requests for every document without an English Abstract file. There were 2.9 usgEn requests for every document with an English Abstract and 2.7 usgEn requests for every document without an English Abstract file. The usgBi data was not large enough to provide conclusive results.

The above figures suggest that having associated English Abstracts with documents did not affect the number of requests for those documents by usgMi clients. It did affect the number of requests for those documents by usgEn clients. UsgEn clients were 1.1 times more likely to request a document if it had an associated English Abstract.

3.5.7 How Long to Read Documents?

Reading of Niupepa documents has been defined as apparent document reading (see §2.5.1). An apparent read time of a document is the time interval from when one document is requested to when another document or page of the interface is requested. We cannot be sure that the client was only reading the document during this time; hence, these times are coined apparent read times.

The average apparent document read times by usgMi clients of the different document types were calculated and are displayed in *Table 3-11*. Also displayed are the average apparent read times by usgEn clients, by usgBi clients and the total average apparent read times.

Average Apparent Read Times of Documents				
(all times in seconds)	usgMi	usgEn	usgBi	All
Average Read of all documents	71.5	55.3	71.7	58.3
Average Read of Full size image	112.0	93.7	120.0	96.9
Average Read of Extracted Text	53.0	32.6	47.8	36.7
Average Read of Preview Image	35.5	26.8	23.9	27.9

Table 3-11: Comparing average apparent read times of document types in usgMi, usgEn and usgBi usage sessions.

The average apparent read times listed in *Table 3-11* suggest that usgMi clients spend, on average, less than two minutes reading the full size images, less than one minute reading the Extracted Text formats and a little over half a minute reading the preview images.

Table 3-11 indicates that all three types of usage clients spend the longest times reading the full size images and the least time reading the Preview Images. The figures also show that the usgMi sessions have longer apparent read times than the usgEn sessions for all types of documents.

3.5.8 Summary—Document Access Behaviours

This section analysed how te reo Māori clients in usage sessions accessed Niupepa documents. Statistics drawn from TLA examined seven document retrieval behaviours exhibited by usgMi clients.

The statistics revealed that using a query was the most favoured method used by usgMi clients to access documents. Using a query was six times more popular than the next most favoured method, browsing the collection by publication listing. Requesting documents in their extracted text form was twice as popular with usgMi clients as requesting documents in their full size image form. UsgMi clients showed a strong preference for accessing documents written in te reo Māori and were the clients least likely to access documents that weren't written in te reo Māori. There were a higher number of requests by usgMi clients for documents that were published in the years: 1842, 1857–1863, 1874–1878, 1898–1905 and 1924–1930. UsgMi clients were more than twice as likely to request a newspaper page if that page happened to have an associated extracted text file. UsgMi clients did not show an increased preference for newspaper pages with associated English Abstracts files. UsgMi clients, on average, required almost two minutes to read a full size image, almost one minute to read an extracted text file and 35 seconds to read a newspaper preview image.

This brief summary concludes the examination into how Niupepa documents were retrieved by clients in te reo Māori. The analysis of how the Niupepa digital library is used by clients in te reo Māori is undertaken in three parts; an analysis of how the interface was used, of how documents were accessed and on how queries were undertaken. Having completed the analysis of how documents were accessed, the next stage is to examine how queries were submitted by usgMi clients. This analysis begins in the following section.

3.6 Query Behaviours

The purpose of this section is to analyse query behaviour of usgMi clients in the Niupepa digital library. Query is the term used to describe the action taken by a client who has entered a query term or terms into the input box of the Niupepa search engine and then clicked the ‘Rapua/Begin Search’ button.

Query statistics were generated by running Perl scripts over the Niupepa web log dataset. There are nine query behaviours analysed in this section:

- the number of queries submitted in usgMi sessions
- the number of query terms used in usgMi queries
- the topics queried by usgMi clients
- the language of query term
- the altering of query parameters in usgMi queries
- the use of query operators in usgMi queries
- the number of documents requested following a usgMi query
- the formats favoured following a usgMi query
- the apparent document read times following a usgMi query.

These nine query behaviours comprise the next nine sub-sections and provide a description of how te reo Māori clients queried the Niupepa digital library.

3.6.1 How Many Queries Were Submitted?

The numbers of queries submitted in each usgMi session were analysed by TLA. There were 975 (57.3%) usgMi sessions that involved at least one query. In comparison, there were 3,317 (67.0%) usgEn sessions and 72 (80.0%) of usgBi sessions that also involved the submission of at least one query. The percentages of queries submitted per usgMi sessions are displayed in *Figure 3-28*. Also displayed are the percentages of queries per usgEn and usgBi sessions.

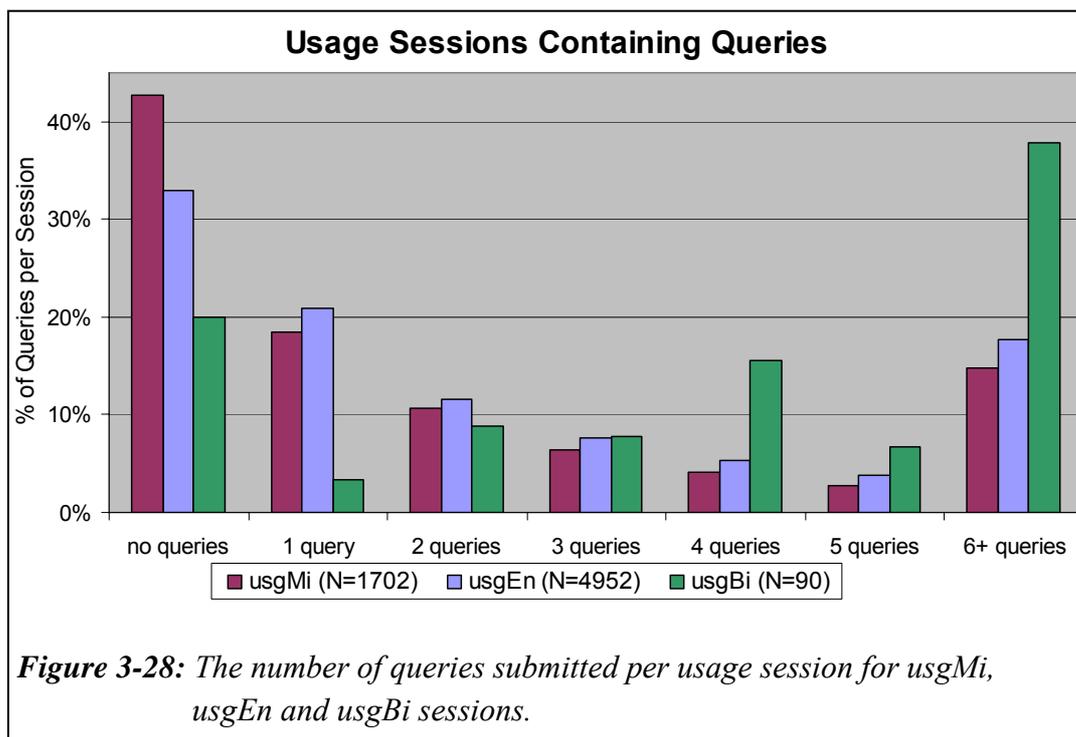


Figure 3-28 shows the usgMi sessions have a stronger presence in the nil to low number of queries submitted while the usgBi sessions have a stronger presence in the medium to high number of queries submitted. The usgEn sessions fall between these two extremes.

It was reported earlier (in §3.5.1) that 77.7% of documents requested by usgMi clients were a result of undertaking a search, which seems to contradict the numbers and figures recorded above. What these numbers do indicate is that usage session queries led to a high number of documents being accessed. The number of documents requested following a query is discussed further in §3.6.7.

3.6.2 How Many Query Terms Were Submitted?

The number of query terms submitted with each usgMi query was examined. Almost half of all usgMi queries consisted of just one query term; just over a quarter consisted of two query terms and the other quarter of usgMi queries

contained three or more query terms. The actual numbers and percentages of query terms per query are listed in *Table 3-12*.

Number of Query Terms Submitted Per Query								
	usgMi		usgEn		usgBi		Total	
Total Queries	4,767	100%	17,227	100%	504	100%	22502	100%
1 term/query	2233	46.8%	8554	49.6%	237	47.0%	11024	49.0%
2 terms/query	1412	29.6%	5475	31.8%	166	32.9%	7053	31.3%
3+ terms/query	1122	23.5%	3202	18.6%	101	20.0%	4425	19.7%

Table 3-12: The number of query terms submitted per query in usgMi, usgEn, usgBi and the total usage sessions.

While the percentages of query terms per query look similar for queries undertaken by usgMi, usgEn and usgBi clients the numbers are in fact significantly different. A chi square analysis on the counts generates a p value of $2.5 * 10^{-13}$ for the usgMi-usgEn comparison and a p value of $9.7 * 10^{-11}$ for the usgMi-usgBi comparison. These figures indicate that the usgMi clients had a statistically significant different preference for the number of terms that were submitted per query than the usgEn and usgBi clients. It can be seen that the usgMi clients submitted the highest number of queries with three or more query terms (23.5%). The usgEn clients submitted the highest number of queries with just the one query term (49.6%). The usgBi clients, by contrast, have the highest percentage (32.9%) for submitting queries which include two query terms.

3.6.3 What Topics Were Queried?

All query terms submitted by clients of the Niupepa digital library were recorded in the web log files. The TLA process collated these terms and grouped them into usage session types. Clients in usgMi sessions submitted 3,379 query terms in 4,766 queries. The term *tioriori* (a figurative term for a warrior) was entered the most often (26 queries) by usgMi clients.

Clients in usgEn sessions submitted 10,797 query terms in 17,227 queries. The term *whakatauki* (proverb) was entered the most often (62 queries) by usgEn clients. Clients in usgBi sessions submitted 322 query terms in 504 queries. The term *arawa* (a tribal name) was entered the most often (13 queries) by usgBi clients.

The top 20 requested query terms by usgMi clients, by usgEn clients and by usgBi clients are listed in *Table 3-13*. Similar terms with different capitalisation or different use of quotation marks are treated as different terms in this list.

The query terms were grouped into topic categories. The fifteen topic categories defined to classify the query terms:

- Names of people—specifically individual people, e.g. *Rewi Maniapoto*
- Iwi (tribes)—includes hapū (sub tribe), nationality and ethnicity e.g. *Tuwharetoa, Chinese*
- Social institutions and roles—e.g. *Kingitanga, kaitiaki*
- Place names—e.g. *Kawhia, Parihaka*
- Historic events and phenomena—e.g. *Treaty of Waitangi, Waikato Wars*
- Cultural activities and concepts—e.g. *haka, whaikorero*
- Cultural phenomena—items of an ethereal nature e.g. *taniwha,*
- Actions and attributes—e.g. *aroha, koingo*
- Geographical features—e.g. *moana, oneone*
- Celestial and seasonal phenomena—e.g. *matariki, christmas*
- Publication and information sources—e.g. *te hokioi, obituaries*
- Plant and animal names—e.g. *apple, tohora*
- Objects—general inanimate objects e.g. *pounamu, rorohiko*
- Language—greetings and particles—e.g. *e hoa, taku*
- Ambiguous—words that cannot be categorised because they have possible referents in several categories e.g. *Te Paki o Matariki* (is this a celestial object, social institution or a publication?)

Query Terms Submitted Most Often		
usgMi	usgEn	usgBi
tioriori (26)	whakatauki (62)	arawa (13)
friend (24)	makaka (53)	parihaka (8)
harakeke (22)	Whakaki (48)	pataua (8)
Tuwhenua (19)	waiata (47)	Ratana (8)
nakahi (14)	awaitaia (40)	"he kupu whakamarama" (5)
rangatiratanga (13)	haka (39)	kingitanga (5)
tuhoe (11)	matariki (32)	ngapuhi whakatauki (5)
mangamaunu (10)	Te wharepu (28)	taipari (5)
Kingitanga (9)	parihaka (27)	waka hokohoko (5)
tutakangahau (9)	treaty (27)	"local government" (4)
pounamu (8)	oruanui (26)	hawaii (4)
rorohiko (8)	motakotako (24)	he kupu whakamarama (4)
taharoa (8)	Tapiri (24)	king (4)
tawhiao (8)	Parihaka (23)	ko au ko ngati whatua (4)
whiro tirea (8)	treaty of waitangi (23)	land disputes (4)
1900 (7)	utiku potaka (23)	ngati koroki ki raukawa (4)
moana (7)	Chinese (20)	proverb (4)
rangatira (7)	pipiwharauoa (19)	tohora (4)
taku (7)	ihaka paehoka (18)	turanga baker (4)
Te Pipiwharauoa (7)	kewene (18)	waikato wars (4)
waewae kura (7)	Mu+aupoko+ (18)	wao tapu nui a tane (4)
Whakaminenga 1835 (7)	Tainguru (18)	
	tawhiao (18)	
	te puke ki hikurangi (18)	

Table 3-13: The query terms (with counts) most commonly submitted by usgMi, usgEn and usgBi clients.

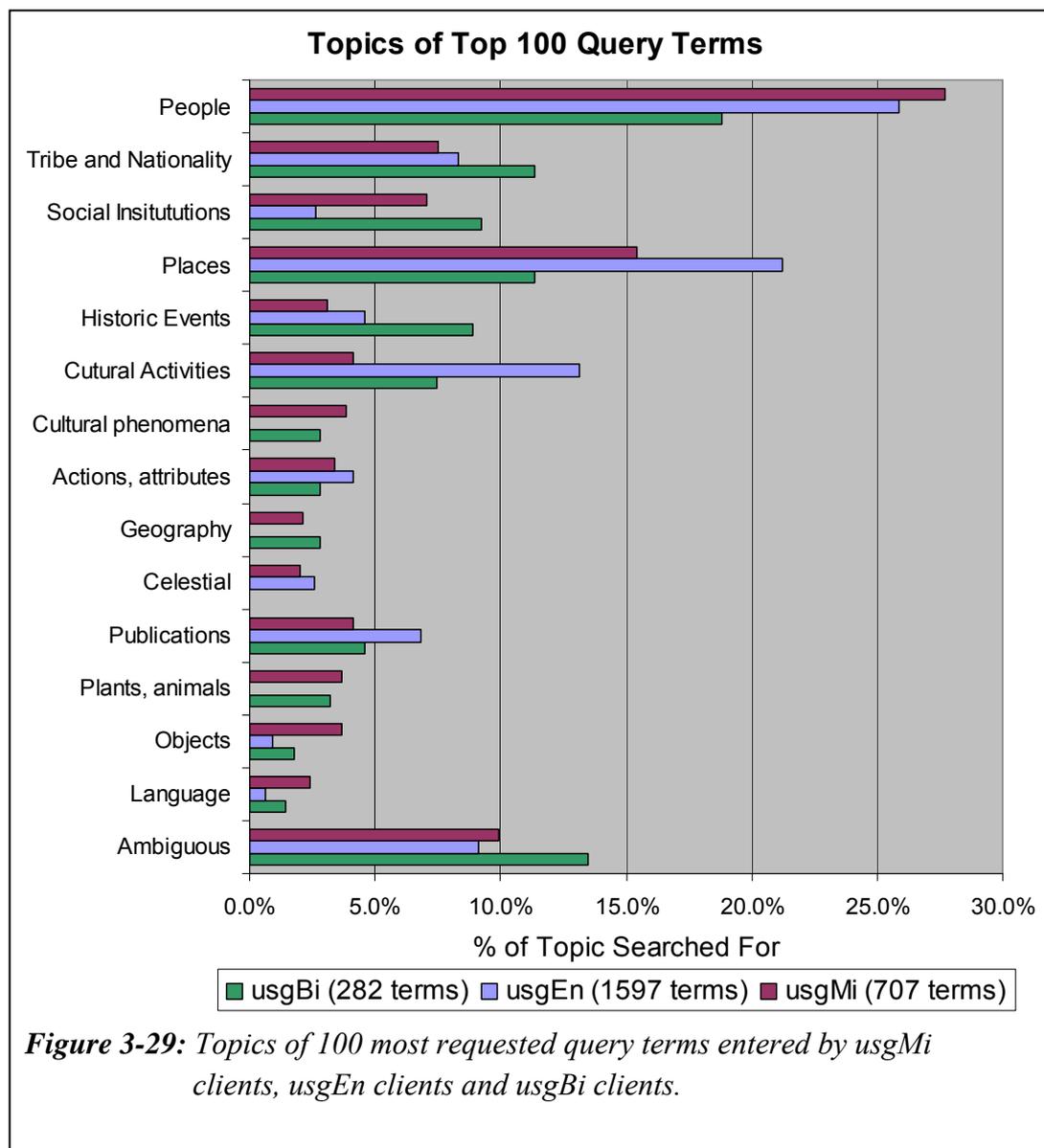
Trying to determine the topic of a query by simply examining the query term is highly speculative; there is no definite way to tell the purpose of a user's searching activity simply by noting the query term that has been submitted.

For example, the most common term submitted by usgMi clients was *tioriori*. There were 26 requests for this term (the next highest was 24). Williams (1985) gives three definitions of the base word, *tiori*: 1) hold up to view, 2) wave to and fro, 3) conspicuous. Williams goes further to explain that *tioriori* is a form of kite, sometime called *manu tioriori*, which is applied to a decoy bird and figuratively to a brave warrior. As the query term has no capitals one could expect that this query term could perhaps be classified as either an attribute or an object. However, when this term is searched in the Niupepa collection 16 documents are found that include this term, 13 of which are names of people and the remaining three are from the phrase *manu tioriori*. One could speculate that some of these results must be useful to the user, hence the high number of searches for this term. Further, the term *tioriori* appears in other queries entered by usgMi clients: *ihaia tioriori* (four requests), *niwha wharerangi tioriori* (three requests), *tioriori kaiwhakawa* (three requests), and *tioriori te haira* (one request), all are queries for an individual. Consequently the term *tioriori* was categorised, albeit somewhat apprehensively, as a query for a person.

Defining the query terms into categories is a laborious process that involves a lot of speculation. Consequently, only the top 100 unique query terms were examined and categorised into the topic headings. For the usgMi sessions this categorisation involved 111 terms that were queried 707 times. For the usgEn sessions this categorisation involved 102 terms that were queried 1,596 times. For the usgBi sessions it involved 100 terms that were queried 282 times. These query terms were categorised into topic categories, the percentage of each term per category was graphed and is displayed in *Figure 3-29*. Although this method of analysis is speculative some general query behaviours are apparent.

The most common topics queried by usgMi clients were names of people and place names. Although 10% of the topic categories were ambiguous usgMi clients did show an interest in all the other topic categories.

Names of people and place names were also the most popular query categories of usgEn and usgBi clients. UsgEn clients also showed a high interest in cultural activities and concepts, publication and information sources and the topic tribes, nationalities and ethnicities. UsgBi clients showed a high number of ambiguous query topics and interests in a wider range of topics.



3.6.4 What Was the Language Of the Query Terms?

The query terms were written in either te reo Māori or English. The content of the Niupepa digital library that was queried consisted not only of the 35,482 newspaper pages but also 313 English Abstracts and 22 Bibliographic Commentaries. Of this content, 53.3% was available only in te reo Māori, 43.9% was available in te reo Māori and English and the remaining 2.8% was available only in English. As there was a large amount of content in te reo Māori, query terms submitted in te reo Māori are likely to yield more results.

The language of the query terms was noted for the top 100 query terms in each usage category and the percentages have been plotted in *Figure 3-30*.

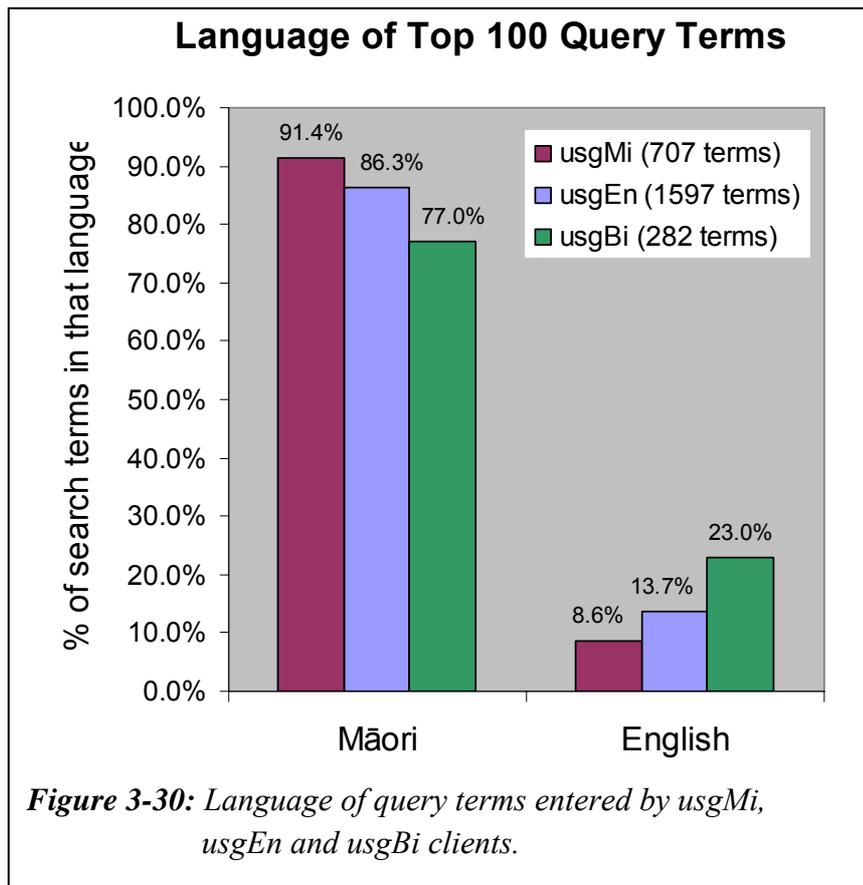


Figure 3-30 shows that 91.4% of the top 100 query terms submitted by usgMi clients were written in te reo Māori. The query terms and their frequencies entered

in English by usgMi clients were: *friend* (24), *maori land court* (6), *death notices* (5), *east coast watch* (5), *Troy* (4), *oil* (4), *prosser* (4) and *Chinese* (4).

The usgEn sessions also had a high percentage of query terms in te reo Māori at 86.3%. This high percentage perhaps suggests that even though these clients are undertaking activities with the interface language set to English they do have an understanding of te reo Māori words. UsgBi sessions have the lowest percentage of terms submitted in te reo Māori at only 77.0%. This lower percentage could suggest that clients defined as usgBi clients were specifically targeting English query terms to assist in the information retrieval process.

3.6.5 Were Query Parameters Altered?

The Niupepa web site offers the user the ability to alter a number of query preferences. These include options to match upper case and lower case characters, to ignore word endings, to retain a list of query terms used, to set the number of results per page and to search on *Some* or *All* of the query terms. Aside from the *Some/All* option, all of these query preferences must be altered on an interface page separate from the Query page, that is the Preferences page (see *Figure 3-2*).

Users seldom altered the query parameters on the Preferences page. Less than 0.13% of usgMi clients altered query parameters on the Preferences page. This behaviour was also displayed by usgEn clients, less than 0.22% altering these preferences, and usgBi clients, less than 0.37% altering these preferences.

However, the one query parameter that was able to be altered on the actual Query page, the *Some/All* option was altered by 10.6 % of the Niupepa digital library clients. The *Some/All* option was least altered by usgMi clients; only 8.4% of the usgMi clients altered this parameter. However, 10.7% of the usgEn clients and a high 23.3% of the usgBi clients chose to alter this parameter when undertaking queries.

3.6.6 Were Query Operators Used?

Operators can be used in queries to broaden or narrow search results. There are two types of operators recognised by the Niupepa digital library search engine; logical operators (and, or, not, *, +, -), and phrase defining operators (quotation marks).

Results from TLA showed that the use of logical operators was low when undertaking a query. Less than 0.9% of usgMi queries used logical operators. The use of logical operators in usgEn queries (0.4%) and usgBi queries was also low (0.5%).

However, there was significant use of the quotation marks to define precise phrase searching. Of the 4,767 usgMi queries submitted 825 (17.3%) involved the use of quotation marks to specify exact phrase matching. This rate was higher than in usgEn queries (6.4%) and usgBi queries (8.3%). It has been noted that usgMi clients entered more query terms when undertaking queries (see §3.6.2); thus it seems probable that usgMi clients would also be more likely to use an operator, such as quotation marks, to assist in defining results in multiple term queries.

3.6.7 Documents Requested Following a Query

The number of documents requested by usgMi clients following the submission of a query was revealed by TLA. There were 4,767 queries submitted in usgMi sessions, directly leading to the request of 16,081 documents; a return rate of 3.4 documents per query. There was also an average of 4.6 documents requested by usgEn clients following a usgEn query and 3.2 documents requested by usgBi clients following a usgBi query.

These figures indicate that usgMi clients (and usgBi clients) did not examine as many search results as usgEn clients. This low request count following a query could be because usgMi (and usgBi) searches were more accurate in returning what the client was looking for or it could simply average that usgMi (and usgBi) clients did not like browsing through the search results pages.

3.6.8 Formats Favoured Following a Query

Statistics produced from the TLA process showed which document formats were requested following the submission of a query. The figures were grouped by language of use, allowing the collation of numbers and percentages of documents requested by usgMi clients following a usgMi query. These percentages are displayed in *Table 3-14*, along with percentages for usgEn and usgBi clients. To assist with a query/non-query comparison, figures and percentages are also listed displaying the document formats requested from activities that did not include a query.

Document Formats Requested following a Query								
	usgMi		usgEn		usgBi		Total	
	Non Query	After Query	Non Query	After Query	Non Query	After Query	Non Query	After Query
Total	6,167	16,081	22,848	79,192	683	1,589	29,698	96,862
Extract. Text	40.9%	69.7%	48.3%	56.8%	51.5%	55.8%	46.8%	59.0%
Full image	50.3%	26.7%	38.0%	38.4%	41.9%	33.4%	40.6%	36.3%
Preview	7.2%	3.5%	12.1%	4.6%	6.1%	10.6%	10.9%	4.5%
Undefined	1.7%	0.2%	1.7%	0.2%	0.4%	0.3%	1.6%	0.2%

Table 3-14: Comparing document formats requested not involving a query with those requested following a query from usgMi, usgEn and usgBi clients.

Table 3-14 reveals that the usgMi clients were much more likely to request an extracted text file after a query than they were if a query was not involved. This characteristic was shown by all client types. Perhaps this increased preference is understandable as the icon links on Niupepa results page all link to the Extracted Text version of the Niupepa documents. However, the increased preference for an extracted text format is much higher in usgMi sessions (28.8%) than in usgEn sessions (8.5%) and in usgBi sessions (4.3%).

3.6.9 Document Read Times Following Query

The apparent read times of documents following a query were compared with the apparent read times of documents when no query was involved. These times, in seconds, are displayed in *Table 3-15*.

(times in seconds)	usgMi		usgEn		usgBi	
	Non Query	After Query	Non Query	After Query	Non Query	After Query
All Documents	87.3	65.8	57.7	54.6	89.2	63.4
Full size Image	126.8	101.8	85.4	96.0	128.1	115.6
Extracted Text	51.4	53.4	44.9	29.7	66.3	40.5
Preview Image	27.1	43.7	27.4	26.3	20.1	25.0

Table 3-15: Comparing apparent read times of non-query documents with apparent read times of query documents by usgMi, usgEn and usgBi clients.

It appears that the usgMi apparent read times of documents were affected by whether or not the document had been requested by a query activity. In particular, the usgMi apparent read times of a full size image were reduced following a query (126.8 to 101.8 seconds) and the usgMi apparent read times of preview images were increased (27.1 to 43.7 seconds) following a query. These times suggest that usgMi clients spent less time reading a full size image after submitting a query and more time reading a Preview Image. Read times of extracted texts remained similar whether a query was submitted or not.

Table 3-15 shows that the usgEn and usgBi apparent read times of extracted texts did change significantly following a query, and smaller changes were also noted in the usgEn and usgBi apparent reads of full size and preview images.

3.6.10 Summary—Query Behaviours

Section 3.6 analysed how te reo Māori clients in usage sessions undertook queries in the Niupepa digital library. Statistics drawn from TLA examined nine query behaviours exhibited by usgMi clients.

The statistics revealed that over 57% of usgMi sessions did involve the submission of a query. Almost half of all usgMi queries consisted of just one query term; just over a quarter consisted of two query terms and the other quarter involved three or more query terms. The most common topics queried by usgMi clients were names of people and places with a small interest also shown in a wide range of other topics. Of the top 100 query terms submitted by usgMi clients, 93% were written in te reo Māori. Query parameters were not altered on the Preferences page (0.1% of usgMi clients) but the Some/All parameter was altered on the Query page by 11% of the usgMi clients. Logical operators were rarely used by usgMi clients (0.9%) but quotation marks were used by 13% to undertake exact phrase matching. For every query submitted, usgMi clients accessed 3.5 documents and were 30% more likely to request the Extracted Text format following a query than they were if a query had not been submitted. After a query was submitted the apparent read times reduced for full size image formats, increased for preview image formats and remained approximately the same for Extracted Text formats.

Having summarised how te reo Māori clients used the Niupepa digital library we are now in a position to deal with two further usage questions: how did usage in te reo Māori differ from usage in English, and how was the Niupepa digital library used by those clients who chose to use both te reo Māori and English? These questions will be addressed in the next two sections.

3.7 Te Reo Māori Usage/English Language Usage

Although the purpose of this chapter is to analyse how te reo Māori was used by clients of the Niupepa digital library, there is also value in detailing how the usage in te reo Māori differed from the usage in English. If differences can be defined then these differences can be incorporated into digital library design and implementation, which could further increase the assistance that digital libraries can give to indigenous languages. The language usage differences have been subdivided into; behaviour differences in expMi and expEn sessions and behaviour differences in usgMi and usgEn sessions.

Behaviour differences in exploratory sessions

Even though the exploratory sessions represented less than 1% of the requests to the Niupepa collection they did highlight three important behaviour differences. First, the expMi sessions were, on average, longer in time (199 versus 136 seconds) and in the number of requests (2.4 versus 2.3 requests) than expEn sessions.

The second behaviour difference noted was that in exploratory sessions different interface pages were requested by expMi and expEn clients. Both usage session types requested the home page as a first preference. However, as a second preference the expMi clients preferred to request the About English Abstracts page, whereas the expEn clients preferred to access the NZDL home page.

The average apparent read time of interface pages was longer for clients who used the interface in te reo Māori. The average expMi apparent read time of interface pages was 2.4 minutes compared with 1.8 minutes for expEn clients.

Behaviour differences in usage sessions

The Niupepa digital library was used more by usgEn clients than by usgMi clients. UsgEn clients were responsible for 79.6% of the total usage requests while usgMi clients were responsible for 18.1% (2.3% were submitted by usgBi clients). On average, usage sessions in English were longer than usage sessions in te reo Māori. The usgEn sessions were longer in time (16.8 compared with 21.3 minutes) than the usgMi sessions and in the average number of requests (18.4 compared with 27.8 requests) per session.

Similarly, the average apparent read time of *interface pages* by clients who used the Niupepa digital library in te reo Māori was longer than for those who used it in English. The average usgMi apparent read time of interface pages was 52 seconds compared with 49 seconds for usgEn clients. A similar pattern occurred in average apparent read times of *documents*; the average usgMi apparent read time of documents was 72 seconds, the average usgEn apparent read time was 55 seconds.

The most popular method of accessing documents was by submitting a query. Queries were not as popular with usgMi clients (77.7%) as with usgEn clients (81.6%). The second most popular method to select documents with usgMi clients was to use the series list while the second most popular method with usgEn clients was to use the date list.

Clients of the Niupepa digital library in te reo Māori (usgMi) preferred to request documents that had a reo Māori component. UsgMi clients displayed a higher preference for requesting te reo Māori newspapers, documents that were written in te reo Māori. Alternatively, usgEn clients had higher preferences for requesting bilingual newspapers, English language newspapers, English Abstracts and Commentary documents (which were written primarily in English). UsgMi clients were twice as likely to access documents with an associated text file as they were to access files without an associated test file. UsgEn clients were almost five times as likely to access documents that had an associated extracted text file. UsgEn clients showed a higher preference for accessing documents with an associated English Abstract whereas usgMi users did not show preferences for documents with an English Abstract.

As mentioned above, submitting queries was less popular with usgMi clients. However, the usgMi clients showed a higher preference for submitting queries with three or more query terms, whereas the usgEn clients also showed a preference for submitting terms with only one or two query terms. Perhaps because of this difference the usgMi clients were three times as likely to use quotation marks surrounding the query terms to activate phrase searching. The usgMi clients were less likely than the usgEn clients to alter the Some/All parameter to further refine queries. Fewer documents were requested by usgMi clients (3.6) following a query than by usgEn clients (4.8).

The submission of a query affected the types of documents that were requested. When a query had not been submitted usgMi clients preferred to request full size images. However, following a query usgMi clients had a much higher preference for requesting extracted text documents, while the preferences for requesting full size images and preview images reduced by almost half. UsgEn clients did not show the same differences following a query. UsgEn clients displayed the highest preference for accessing an extracted text document whether a query was submitted or not. The only difference following a query by usgEn clients was a reduction in the preference for accessing the preview images, which subsequently led to an increase in the preference for accessing the extracted text documents.

The submission of a query also affected the apparent read times of documents. For usgMi sessions, the apparent read times following a query decreased for full size images, remained the same for extracted text documents but increased for preview images. For usgEn sessions, the apparent read times increased following a query, increased for full size images, decreased for extracted text documents and stayed the same for preview images.

A reo Māori to English interface language switch is something that usgEn showed a strong preference for. When usgEn clients undertook this switch it usually occurred at the start of a session. When usgMi clients switched the interface language from reo Māori to English they did it towards the end of a session. Both user groups switched from English to te reo Māori, although not to the same extent and the switches were mostly occurred at the beginning or end of a session.

3.8 Bilingual Usage

The decision to divide the usage sessions into three types of language users, usgMi clients, usgEn clients and usgBi clients, was based on earlier research (T. T. Keegan & Cunningham, 2005a). This research suggested there were three distinct types of clients; those that undertook sessions primarily with the interface language set to te reo Māori (usgMi), those that undertook sessions primarily with the interface language set to English (usgEn), and those that undertook sessions with the interface language set to both te reo Māori and English for a significant portion of the session (usgBi).

Web log analysis undertaken in this chapter has suggested that these usgBi sessions were not the result of a separate and distinctive user group but, instead, were the result of a unique activity that was undertaken by a small number of clients. Only 1% of all usage sessions were classified as usgBi sessions. Of the 90 usgBi sessions, none was a return session, suggesting this is an activity that is undertaken only once by a client. At least 13 of the clients that undertook this activity also undertook activity as a usgEn client (11) or as a usgMi client (2).

The requirements for usgBi sessions ensured that these sessions would be long sessions. UsgBi sessions were defined as having at least 20% of the requests in both te reo Māori and English, at least ten requests and at least three language switches. Consequently, usgBi sessions were, on average, twice as long in duration as usgMi and usgEn sessions, and consisted of twice as many requests.

UsgBi clients requested the Preferences page almost twice as often as usgMi and usgEn clients. Aside from the home page, the Preferences page is the page on which the language of the interface is switched so it seems likely that usgBi clients were accessing this page more, as all usgBi sessions had at least three language switches. The average apparent reading time of interface pages was longer for usgBi sessions, again, an attribute possibly caused by the longer time it takes to switch the interface language through the Preferences page.

Three query activities were undertaken differently by usgBi clients; the number of queries submitted per session by usgBi clients was higher than by usgMi or usgEn clients, the number of English terms submitted per query was the highest in usgBi sessions, and the Some/All query parameter was altered by usgBi clients at least twice as often as by usgMi or usgEn clients. All three of these activities suggest that the usgBi sessions were sessions where submitting a query was a primary activity, and that both te reo Māori and the English language were used together in a creative way to assist with this querying activity. The usgBi clients accessed fewer documents per query than usgMi or usgEn clients, suggesting that perhaps these queries were readily returning the desired result.

Many of the other usage behaviours exhibited by usgBi clients were similar to those recorded for usgMi and usgEn clients. The apparent reading times of documents, the type of documents requested, the number of query terms, the topic of query terms, the activities following a query and the positions in the session where the interface language was switched were similar to, or between the extremes of, usgMi and usgEn behaviours.

These results suggest that te reo Māori, in conjunction with the English language, was used as a deliberate strategy to assist in information retrieval activities by bilingual clients. Bilingual clients have the ability to use the Niupepa digital library in te reo Māori or English. It appears that a small number of clients were choosing not to use it in one language or the other but, rather, were choosing to use it in *both* languages in the one session. This use of two languages appears to be an effective information retrieval strategy and was used by a small number of clients to request a relatively high number of documents.

3.9 Summary

The purpose of this chapter has been to use Transaction Log Analysis (TLA) to determine how te reo Māori was used by clients of the Niupepa digital library in 2005. This chapter began with a full description of the Niupepa interface screens and options available to clients.

TLA was used to filter and categorise all requests to the Niupepa digital library into one of three types of sessions; single request sessions, exploratory sessions, and usage sessions. By analysing this data it was possible to answer the question, was te reo Māori used by clients of the Niupepa digital library? Results from the analysis showed that te reo Māori was used by Niupepa clients in 20.5% of all requests. Te reo Māori was used in 35% of single request sessions, in 46% of the requests in exploratory sessions and in 19% of the requests in usage sessions.

Exploratory session analysis was undertaken to determine how clients used te reo Māori to explore the Niupepa web site but then not go on to use it. The 180 exploratory sessions only represented a small number of requests to the Niupepa digital library (less than 1%). These types of clients, expMi clients, only made two or three requests and mostly completed the session in less than half a minute. The Niupepa Home page was the interface page mostly requested by expMi clients, although the About English Abstracts page was also significantly requested.

As opposed to the limited amount of information available from exploratory session analysis, usage session analysis was able to provide more extensive information on how te reo Māori was used by Niupepa clients. Most of the requests (94%) to the Niupepa digital library were requests classified as usage session requests. Data on usage sessions showed that sessions undertaken by usgMi clients averaged 30 minutes and 18 requests. Only 6% of sessions were return sessions. The usage analysis was collated into three categories; interface usage by usgMi clients, document access by usgMi clients and queries submitted by usgMi clients.

Interface usage analysis by usgMi clients showed the Niupepa Home page was the interface page most requested, although the Preferences page was also significantly requested. The apparent read time of interface pages by usgMi clients was 52 seconds. The language of the interface was not switched in many usgMi sessions (only 8%); when it was switched it was likely to be from English to Māori and generally in the first half of the session.

Document access analysis by usgMi clients revealed using a query was the most favoured method used to access documents. Overall, requesting documents in the extracted text format was twice as popular with usgMi clients as requesting documents in the full size image format. UsgMi clients showed a high preference for documents written in te reo Māori and low preference for documents written in English. Documents with associated extracted text files were requested twice as much by usgMi clients. The average apparent read times by usgMi clients for documents were almost two minutes for a full size image, almost one minute for an extracted text file and 35 seconds for a preview image.

Analysis of queries submitted by usgMi clients showed that almost 43% of usgMi sessions did not involve the submission of a query. However when a query was submitted almost half consisted of just one query term, just over a quarter consisted of two query terms and the other quarter involved three or more query terms. Names of people and place names were the most common topics queried by usgMi clients. Most query terms (93%) were written in te reo Māori. Query parameters and logical operators were not used by usgMi clients when undertaking queries, aside from the Some/All parameter on the Query page and the speech quote operators to define phrase searching. On average usgMi clients accessed 3.5 documents following a query. Full size images were the document formats favoured by usgMi clients when a query had not been submitted; however post-query the extracted text format was favoured. Apparent read times of documents following a query increased for preview images, reduced for full size images and did not alter for Extracted Text formats.

Activities undertaken on the Niupepa digital library in te reo Māori were compared with activities undertaken in English. Differences noted in the statistics

showed that expMi sessions were longer than expEn sessions but usgMi sessions were shorter than usgEn sessions. Clients in te reo Māori appeared to take longer to read the interface pages and Niupepa documents. Querying was undertaken less by usgMi clients but when it was undertaken it often involved the use of more query terms and fewer results were examined following a query.

The final question raised in this chapter asked how te reo Māori had been used with the English language to assist in information retrieval. A small number of bilingual clients chose to use te reo Māori *and* English in the one usage session. These sessions, usgBi sessions, were longer sessions, concentrating on querying activities, and contained the highest mix of reo Māori/English query terms. UsgBi clients requested the lowest number of documents following a query, an average of 3.2; lower than usgMi clients who requested 3.4 documents and lower than usgEn clients who requested 4.6 documents following a query. These figures suggest that using te reo Māori with English was a strategy that led to more accurate query activity.

The information presented in this chapter has charted te reo Māori usage by clients of the Niupepa Digital library in 2005. It has distinctly shown how te reo Māori has been used by clients to undertake information retrieval behaviour in the Niupepa digital library. This reo Māori use has occurred without prompting or coercion, clients have chosen to use te reo Māori in a modern electronic medium.

The next question to be considered is whether the usage of the Niupepa digital library by reo Māori clients differs between those clients from Aotearoa and those clients who access the Niupepa digital library from overseas. The following chapter uses statistics generated from TLA to undertake a comparative analysis of international and national usage of the Niupepa digital library by clients who undertake their activities in te reo Māori.

4 Comparing International and National Reo Māori Usage

This chapter analyses how te reo Māori was used by international (non-Aotearoa) and national (Aotearoa) clients of the Niupepa digital library. Was te reo Māori used by international clients? If it was, was te reo Māori used in the same way by international clients as it was by national clients?

Statistics generated from Transaction Log Analysis (TLA) were used to compare te reo Māori usage by international and national Niupepa clients in 2005. A comparative analysis examines exploratory and usage activities of international and national clients. Usage activities examined include interface usage behaviours, document access behaviours and query behaviours. The international usgMi clients are shown to be distinct users with their own peculiar information retrieval behaviours.

The method used to differentiate between national and international requests has been described in §2.2.2. Two datasets were available because two mirror web servers served web requests to the Niupepa digital library in 2005. A server called Puka, located in Hamilton, Aotearoa, was responsible for serving all national requests. A server called Thor, in Lethbridge, Canada, was responsible for serving all international requests. Both servers independently logged the requests that they served into web log files. The Puka web file provided raw data about national requests. The Thor web file provided raw data about international requests.

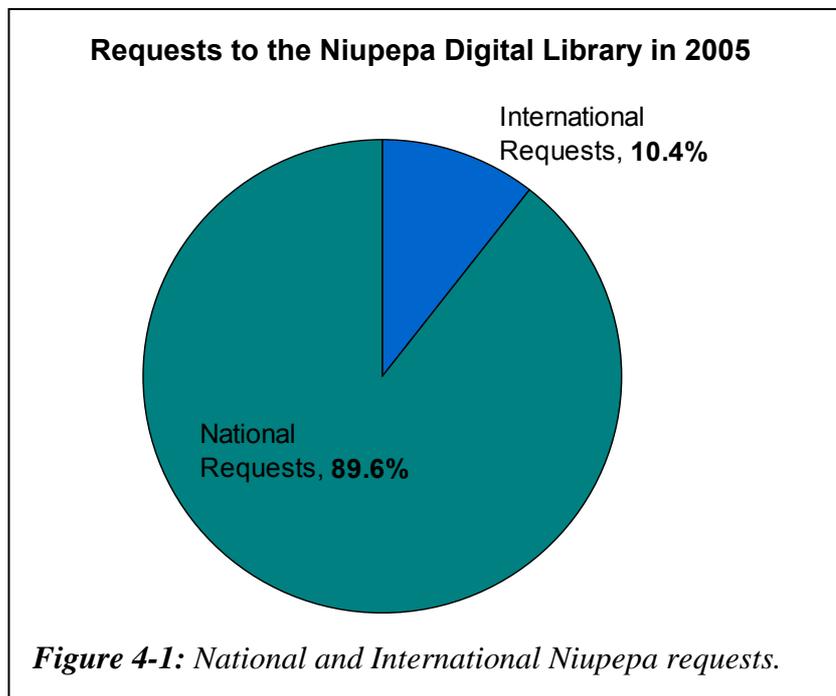
4.1 Was Te Reo Māori Used by International Clients?

This chapter details a comparative analysis of te reo Māori usage of the Niupepa digital library by national and international clients. However before this comparison can begin it must first be established that the Niupepa digital library was used by international clients and then it must be established that it was used by international clients in te reo Māori. This section uses statistics produced from TLA to answer these questions.

4.1.1 Did International Clients Use the Niupepa?

The Niupepa digital library is made available through the New Zealand Digital Library (NZDL) web site (see www.nzdl.org). The NZDL web site received a total of 1,133,543 raw requests in 2005 (see *Table 2-2*) of which 842,403 requests (74.3%) were logged as international requests and 291,140 requests (25.7%) were logged as national requests. However, most of the international requests were requests to other NZDL digital library collections, not the Niupepa digital library. There were only 28,501 international requests (3.4% of international requests) to the Niupepa digital library. In contrast, most of the national NZDL requests were to the Niupepa digital library and not to other NZDL digital libraries. There were 175,864 national requests (60.4% of national requests) to the Niupepa digital library.

Requests described as extraneous requests (see §2.3.2) were removed from the two sets of logs leaving 19,303 filtered international requests and 165,698 filtered national requests. These requests are displayed graphically in *Figure 4-1*.



Although the 19,303 international Niupepa requests are comparatively small in number (10.4%) compared with the number of national Niupepa requests there are still sufficient to undertake meaningful TLA.

When removing extraneous requests it was noted that the international requests had a much higher proportion of requests removed because of web robot activity. There were 9,122 (32.0%) web robot requests removed from the international Niupepa log. There were 8,055 (4.6%) web robot requests removed from the national Niupepa log. This indicates that although some web robot activity did originate in Aotearoa it was almost eight times more likely to have originated from outside Aotearoa.

International clients did use the Niupepa digital library to submit almost 20,000 requests.

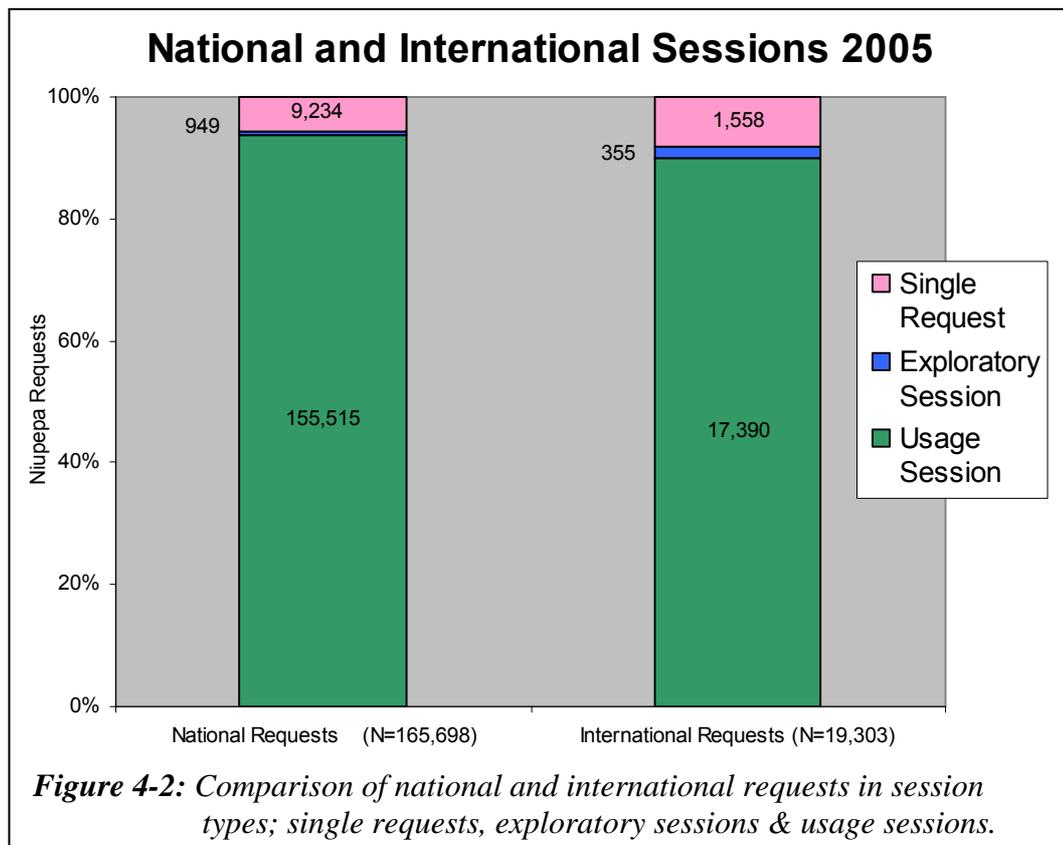
4.1.2 Did International Clients Show Interest in the Niupepa Digital Library?

The proportion of session types can give an indication of how interested prospective clients were in the Niupepa digital library. A high proportion of usage sessions combined with a low proportion of exploratory and single request sessions indicated a high level of interest.

TLA was used to separate the international and national requests into session types. The numbers and percentages of session types are displayed in *Figure 4-2*.

The 19,303 international requests consisted of 1,558 (8.1%) single requests, 355 (1.8%) exploratory session requests and 17,390 (90.1%) usage session requests. The 165,698 national requests consisted of 9,234 (5.6%) single requests, 949 (0.6%) exploratory session requests and 155,515 (93.9%) usage session requests.

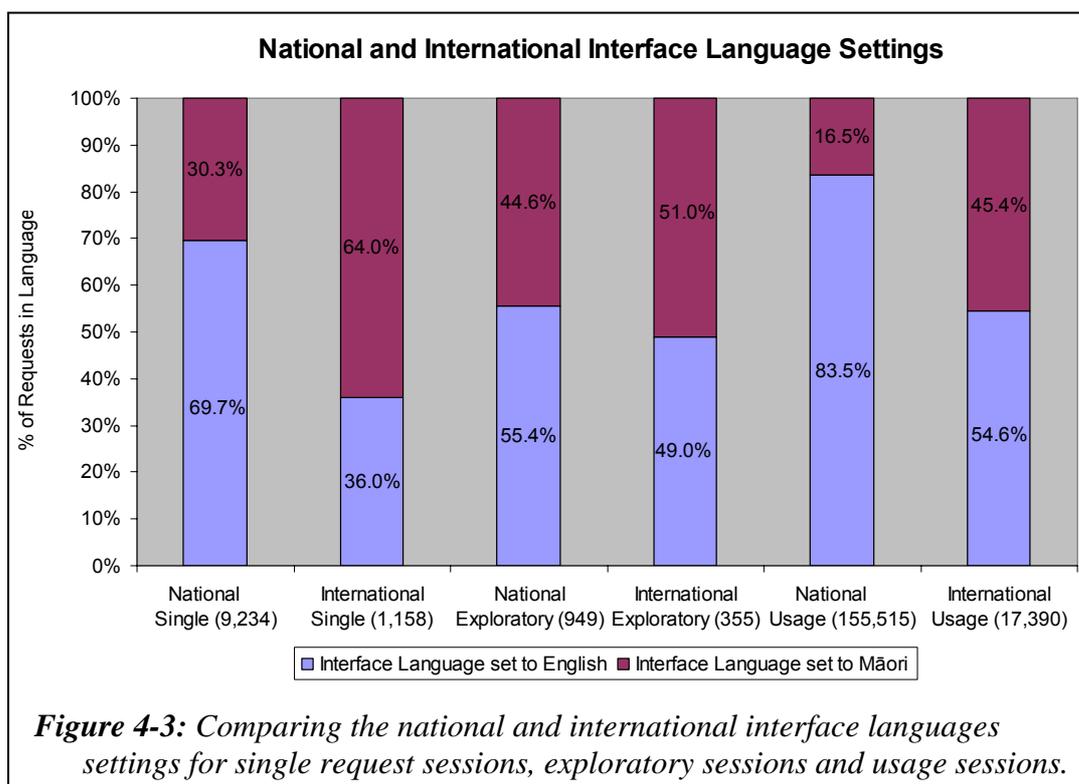
The international requests had a significantly higher proportion of single requests and exploratory session requests. This higher proportion indicates that international clients, having requested a Niupepa interface page, were less likely to then go on and use the Niupepa digital library. As the Niupepa collection presents historic information about Aotearoa it seems reasonable that this information is less interesting to international clients.



4.1.3 Did International Clients Use Te Reo Māori?

Having separated the requests into types of sessions the TLA process was then able to count how many requests were submitted with the Niupepa interface set to te reo Māori and how many requests were submitted with the Niupepa interface set to English. The counts of interface language settings in international sessions and interface language settings in national sessions are displayed in *Figure 4-3*.

Figure 4-3 shows that te reo Māori was used proportionally more often by international clients than it was by national clients. This higher preference for te reo Māori was shown by international clients in all session types.



Usage sessions are the most important to evaluate as they comprise the greatest proportion and give insights into how clients actually used the Niupepa interface. Only 16.5% of national usage requests were undertaken with the interface language set to te reo Māori. However, almost three times that proportion, 45.4%, of international usage requests are requests undertaken with the interface language set to te reo Māori.

International clients did use the Niupepa digital library in te reo Māori and they were almost three times more likely to use it in te reo Māori than their national counterparts.

4.1.4 When Was Te Reo Māori Used?

It is difficult to determine when international requests were submitted from a chronological perspective. Users may log onto the Niupepa web site from any time zone in the world; however the request time and date recorded in the web

logs is the local time in the time zone where the web server is located. The local time of the web server in Aotearoa will not match the local time of international clients. Consequently, international ‘hour of day’ and ‘day of week’ statistics cannot be generated with any degree of certainty. However, the number of requests per month is not affected to the same degree. The months in which international clients submit requests will be recorded correctly in the web log files for at least 97% of the requests but this validity is probably closer to 99%. This accuracy allows a comparison to be undertaken for requests submitted per month by international and national clients.

Percentages of both international and national usage requests in te reo Māori per month were calculated and are displayed in *Figure 4-4*.

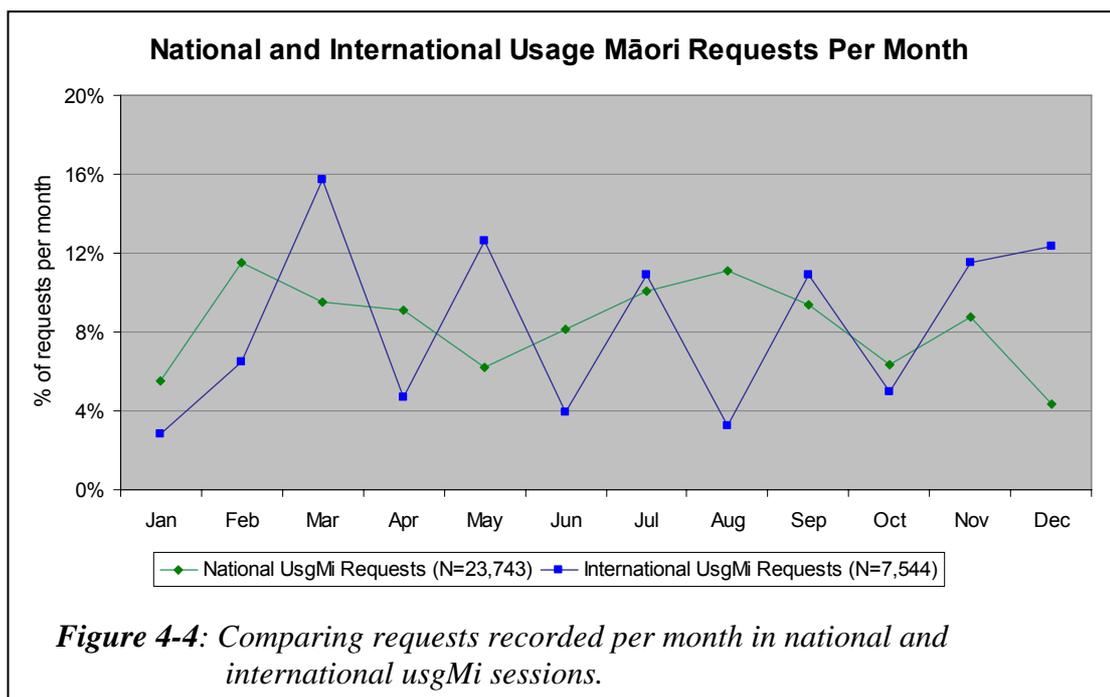


Figure 4-4 clearly shows that the international usgMi requests have monthly cycles of peaks and troughs, whereas the national usgMi requests were more evenly distributed across the year. This raises the question why were the international clients three more likely to undertake requests in te reo Māori in alternative months? The answer can be attributed to the default language setting of the Niupepa web site. In the months of January, March, May, July, September and

November the default language of the Niupepa web site was set to te reo Māori. In the other months the default language was set to English. Aside from the first and last months, the months when the default language was set to te reo Māori are the months during which international clients were three times more likely to use the interface in te reo Māori.

The default language of the web site appears to have had significant impact on which language the clients, in particular international clients, chose to navigate the web site. An extensive examination of this influence is undertaken in Chapter 5.

4.1.5 Summary—Was Te Reo Māori Used?

This section has used statistics generated from TLA to determine if te reo Māori was used by international clients of the Niupepa digital library. First it was established that the Niupepa digital library received 19,303 filtered usage requests from international clients in 2005. International requests comprised approximately 1 in 10 of all usage requests received by the Niupepa digital library. There were a higher proportion of single request sessions and exploratory sessions in international sessions than in national sessions, suggesting that prospective international clients were less likely to use the Niupepa digital library than prospective national clients. However, the international clients were more likely to use the Niupepa digital library in te reo Māori. International clients were almost three times more likely to use the Niupepa digital library in te reo Māori than national clients and were most likely to use the Niupepa interface in te reo Māori when the default language of the interface was set to te reo Māori.

Having established that the Niupepa digital library was used in te reo Māori by international clients, the next stage is to determine how they used te reo Māori and if this usage was different from the usage of te reo Māori by national clients. An analysis of te reo Māori usage by international clients and a comparative analysis of national/international reo Māori usage are detailed in the next four sections.

4.2 International and National Exploratory Behaviours

Clients that requested pages of the Niupepa interface but then did not go on to submit queries or request any documents have been defined as exploratory clients. Requests from exploratory clients comprised less than 1% of the total requests to the Niupepa digital library in 2005. Although this is a low percentage of requests, the data can still be used to compare how national and international clients in te reo Māori explored the Niupepa digital library.

Approximately one quarter of exploratory requests were submitted by international clients and three quarters by national clients. There were 355 requests from 163 exploratory sessions recorded as international requests in 2005. There were 949 requests from 404 exploratory sessions recorded as national requests in 2005. A high proportion of exploratory sessions lasted less than two minutes; 77.5% of national sessions and 84.7% of international sessions. A high proportion of exploratory sessions consisted of just two requests; 79.0% of national sessions and 87.7% of international sessions.

The international and national exploratory sessions were further subdivided by language of use. There were:

- 95 requests in 43 international expMi sessions
- 331 requests in 137 national expMi sessions
- 260 requests in 120 international expEn sessions
- 618 requests in 267 national expEn sessions

Exploratory sessions were categorised as either expMi or expEn. However, they often consisted of just two requests, both for the home page, one with the interface set to te reo Māori and one with the interface set to English. By definition (see §2.4.3), if the exploratory session consisted of just two requests, and the last request was with the interface set to te reo Māori then the session was defined as a usgMi session. Alternatively, if the exploratory session consisted of just two

requests, and the last request was with the interface set to English then the session was defined as a usgEn session. A listing of the international and national exploratory sessions, including which language the interface was set to when requests were submitted, shows:

- 43 international expMi sessions; 79 requests in te reo Māori,
16 requests in English,
- 137 national expMi sessions; 267 requests in te reo Māori,
64 requests in English,
- 120 international expEn sessions; 158 requests in English,
102 requests in te reo Māori,
- 267 national expEn sessions; 462 requests in English,
156 requests in te reo Māori.

The figures listed above indicate that many of the exploratory sessions involved a language switch. TLA was used to examine three further characteristics of exploratory sessions in te reo Māori—the length of expMi sessions, the apparent reading times in expMi sessions and the interface pages requested in expMi sessions. Summaries of these three exploratory characteristics follow.

4.2.1 How Long Were Exploratory Sessions?

Statistics generated by TLA revealed the average lengths of international and national exploratory sessions. The average lengths in time and requests were:

- 3.5 minutes and 2.2 requests in international expMi sessions
- 2.8 minutes and 2.4 requests in national expMi sessions
- 45 seconds and 2.2 requests in international expEn sessions
- 2.5 minutes and 2.3 requests in national expEn sessions

The times listed above indicate that international expMi clients took the longest time, on average 3.5 minutes, to determine that they did not want to use the Niupepa digital library. The international expEn clients were fastest at deciding not to use the Niupepa digital library (average 45 seconds). Just over two requests was the average for all national and international exploratory sessions.

4.2.2 How Long to Read Interface Pages?

The exploratory categories with the longest session lengths were also the categories that exhibited the longest apparent read times of interface pages. The average apparent interface page read times in international and national exploratory sessions were:

- 185.3 seconds in international expMi sessions
- 135.0 seconds in national expMi sessions
- 54.7 seconds in international expEn sessions
- 127.7 seconds in national expEn sessions

The international expMi clients exhibited the longest average read time of interface pages (185.3 seconds). The international expEn clients exhibited the shortest average read time of interface pages (54.7 seconds).

4.2.3 Which Interface Pages Were Requested?

An interface page is a page of the Niupepa web site; it does not refer to a newspaper document in the Niupepa collection. The interface pages requested by international and national expMi clients and international and national expEn clients are listed in *Table 4-1*.

Interface Pages Accessed in Exploratory Sessions								
	expMi				expEn			
	International		National		International		National	
Total pages	95	100.0%	331	100.0%	260	100.0%	618	100.00%
home page	87	91.6%	271	81.9%	243	93.5%	485	78.50%
NZDL home page	0	0.0%	9	2.7%	2	0.8%	58	9.40%
Eng. Abstr. page	7	7.4%	31	9.4%	9	3.5%	26	4.20%
help page	0	0.0%	5	1.5%	2	0.8%	22	3.60%
preferences page	1	1.1%	14	4.2%	4	1.5%	25	4.00%
greenstone page	0	0.0%	0	0.0%	0	0.0%	1	0.20%
unknown page	0	0.0%	1	0.3%	0	0.0%	1	0.20%

Table 4-1: Number and percentages of web site interface pages requested in international and national exploratory sessions.

It can be seen that the majority of the requests, approximately 80% of total national and 92% of total international, were requests for the Niupepa home page. These requests consisted of requests for the home page in te reo Māori, requests for the home page in English and requests for hyperlinks to texts located lower down on the home page. International usgMi clients had a higher preference for requesting the Niupepa home page and, aside from the About English Abstracts page, did not request other Niupepa interface pages.

4.2.4 Summary—Exploratory Sessions

Exploratory sessions in general were short, many simply involving a request for the home page in one language and then a request for the same page in the other language. International expMi sessions were on average slightly longer (3.5 minutes compared with 2.8 minutes) than national expMi sessions and the apparent read times of interface pages were longer (185 seconds compared with 135 seconds). Most of this time was taken reading the Niupepa home page. International usgMi clients did not show the same interest in other pages of the Niupepa interface as national usgMi clients.

4.3 International and National Interface Behaviours

This section analyses how the Niupepa interface was *used* by international reo Māori clients and compares this usage with usage by national reo Māori clients. The purpose of this section is to determine if international clients used the Niupepa interface in te reo Māori differently from national clients.

By definition (see §2.4.2), a usage session is a series of two or more requests from an individual user where each request is no longer than an hour apart and a query is submitted and/or an actual document of the Niupepa Collection is requested. Usage session data is the most useful of all the data analysed as it gives the best insights into how actual users were using the Niupepa interface.

There were 172,905 usage requests recorded in 6,714 sessions for the Niupepa collection in 2005. Of these, 17,390 (10.1%) were recorded as being international usage requests and 155,515 (89.9%) as national usage requests.

The usage sessions requests were categorised into language session types. The counts for the various international and national usage session types were:

- 380 international usgMi sessions consisting of 7,544 requests
- 1,322 national usgMi sessions consisting of 23,743 requests
- 520 international usgEn sessions consisting of 9,340 requests
- 4,432 national usgEn sessions consisting of 128,357 requests
- 14 international usgBi sessions consisting of 506 requests
- 76 national usgBi sessions consisting of 3,415 requests

These figures show that the usage sessions in reo Māori (usgMi sessions) had the highest ratio of international and national requests. Almost a quarter of usgMi requests were from international clients, compared with 1 in 15 of usgEn requests and 1 in 8 of usgBi requests.

Five interface behaviours were examined to compare the international and national usage in te reo Māori of the Niupepa interface. The interface behaviours compared for international and national usgMi sessions were:

- the length of sessions
- the apparent read times of interface pages
- the number of return sessions
- the interface pages requested
- the frequency of interface language switching.

These five interface behaviours are described in the next five sub-sections.

4.3.1 How Long Were Usage Sessions?

Two measurements were used to define length of usage sessions, duration of time (minutes) and number of requests. Statistics generated from TLA were used to compare national and international usage session lengths of the Niupepa digital library in 2005. These comparisons are discussed for both measurements.

How Long in Time Were Usage Sessions?

The durations (in minutes) of the national usgMi sessions were compared with the durations of the international usgMi sessions. The durations of the usage sessions were grouped into time intervals of less than 1 minute, 1–2 minutes, 3–9 minutes, 10–29 minutes, 30–59 minutes and 60+ minutes. The count of sessions that fitted into these categories was collated for national and international sessions. Chi square analysis was undertaken on these counts. The counts and the p values of the chi square analysis are shown in *Table 4-2*.

International and National Duration (in Minutes) of Usage Sessions						
Minutes per session	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Sessions	380	1322	520	4432	14	76
<1	82	374	131	903	0	0
1–2	101	227	146	905	0	4
3–9	89	226	102	798	2	9
10–29	80	234	70	848	4	23
30–59	17	151	47	506	7	25
60+	11	110	24	472	1	15
p values	2.8 * 10 ⁻¹⁰		8.9 * 10 ⁻⁹		Not applicable	

Table 4-2: International and national comparisons for duration groupings of usgMi, usgEn and usgBi sessions showing chi square p values.

A p value of $2.8 * 10^{-10}$ indicates a significant statistical difference in national and international usgMi session durations. Similarly, a p value of $8.9 * 10^{-9}$ indicates a significant statistical difference in national and international usgEn session durations. The low number of usgBi sessions does not give accurate chi square analysis figures so this calculation was not used for usgBi sessions.

Summary statistics were calculated to highlight the international and national duration differences of usgMi, usgEn and usgBi sessions. These summary statistics are displayed in *Table 4-3*.

Table 4-3 indicates that international usgMi sessions on average were completed in almost half the time of national usgMi sessions (10.7 minutes compared with 18.1 minutes). This difference in session duration is statistically significant, as confirmed by the p value displayed in *Table 4-2* ($2.8 * 10^{-10}$). *Table 4-3* also shows that international usgEn sessions were, on average, shorter than national usgEn sessions (13.4 minutes compared with 21.7 minutes) and international usgBi sessions averaged shorter session durations than national usgBi sessions (30.7 compared with 46.0 minutes).

Summary International and National Duration (Minutes) of Usage Session						
Minutes	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Average	10.7	18.1	13.4	21.7	30.7	46.0
Standard Deviation	20.0	32.6	27.2	37.0	17.0	50.3
Most Common	0.0	0.0	0.0	0	11.0	21.0
Least	0	0	0	0	4	3
Lower Quartile	1	0	0	1	16	17
Median	4	5	3	6	33.5	34
Upper Quartile	12	22	13	26	40	52
Most	194	333	239	356	68	256

Table 4-3: International and national comparisons of statistical summaries for durations (in minutes) of usgMi, usgEn and usgBi sessions.

How Long in Number of Requests?

The lengths of usage sessions in requests were also compared between international and national usgMi sessions. The counts of requests per session were grouped into the following intervals 2 requests, 3–4 requests, 5–8 requests, 9–19 requests, 20–39 requests and 40+ requests. Chi square analysis was undertaken on these counts. The counts and the p values of the chi square analysis are displayed in *Table 4-4*.

A p value of 5.1×10^{-4} indicates a significant statistical difference in the count of requests in national and international usgMi sessions. Similarly, a p value of 3.6×10^{-13} indicates a significant statistical difference in the count of requests in national and international usgEn sessions. The low number of usgBi sessions would not give accurate chi square analysis figures so this calculation was not undertaken for usgBi sessions.

International and National Length in Requests of Usage Sessions						
Requests per session	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Sessions	380	1322	520	4432	14	76
2	36	205	51	367	n/a	n/a
3–4	31	185	68	451	n/a	n/a
5–8	90	267	147	780	n/a	n/a
9–19	107	337	140	1179	4	16
20–39	62	185	63	741	5	30
40+	54	143	51	914	5	30
p values	$5.1 * 10^{-4}$		$3.6 * 10^{-13}$		Not applicable	

Table 4-4: International and national comparisons of request groupings in usgMi, usgEn and usgBi sessions with chi square p values.

Summary statistics were calculated to determine exactly what the differences were in international and national usgMi sessions lengths (in request count). These summary statistics are displayed in *Table 4-5*.

International and National Summary of Session Lengths (Requests)						
Requests per session	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Average	19.9	18.0	18.0	29.0	36.1	44.9
Standard Deviation	23.6	27.0	27.5	43.4	21.3	37.8
Most Common	2.0	2	2.0	2.0	10.0	32.0
Least Requests	2	2	2	2	10	13.0
Lower Quartile	6	4	5	6	20	22
Median	11	9	8	13	31	31.5
Upper Quartile	24	20	17	34	53	55
Most requests	192	245	254	598	74	239

Table 4-5: International and national comparisons of statistical summaries for length (in requests) of usgMi, usgEn and usgBi usage sessions.

Table 4-5 indicates that international usgMi sessions averaged more requests (19.9) than national usgMi sessions (18.0). Although this does not appear to be a great difference, the p value displayed in *Table 4-4* ($5.1 * 10^{-4}$) indicates that the difference is statistically significant.

UsgEn sessions and usgBi sessions were different from usgMi sessions in that the average lengths of sessions by request count was longer for national than international clients.

UsgBi sessions were twice as long as the other sessions and had twice as many requests. One of the criteria for defining a usgBi session is that the sessions must consist of at least 10 requests, the result of which means that only long sessions are defined in this category.

4.3.2 How Long to Read Interface Pages?

An apparent read time of interface page is the length of time from when one page of the interface is requested until a subsequent page or document is requested. It is called an apparent read time because we are unable to be certain that the user was *only* engaged with the interface during this time. Statistical summaries of apparent read times in international and national usgMi sessions are displayed in *Table 4-6*. Also listed are apparent read times in international and national usgEn sessions and in international and national usgBi sessions.

It appears that the international usgMi clients took half as long to read a page as the national usgMi clients. A chi square statistical analysis comparing national and international usgMi apparent read times returns a p value of $4.2 * 10^{-32}$. This value implies statistical significant difference.

Apparent Read Times of Interface Pages in Usage Sessions						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Count	387	952	713	5097	93	448
Average (seconds)	27.3	61.4	58.1	47.4	74.6	71.3
Standard Deviation	104.6	179.4	212.4	183.4	228.9	221.8
Mode (seconds)	4	7	4	4	4	4
Shortest (seconds)	0	0	0	0	3	0
Lower Quartile (seconds)	4	6	6	6	7	6
Median (seconds)	5	14	13	12	18	14
Upper Quartile (seconds)	10	40	33	32	39	40
Longest (seconds)	1156	2131	2778	3237	1871	2367
p value	$4.2 * 10^{-32}$		0.61		0.57	

Table 4-6: International and national statistical comparisons of apparent read times of Niupepa interface pages in usgMi, usgEn and usgBi sessions.

While the international and national usgEn apparent read times appear different, with averages of 47.4 and 58.1 seconds, the p value of 0.61 suggests that these are not significantly different. The international and national usgBi sessions, with a p value of 0.57, also have apparent read times that are not significantly different. These p values indicate that if clients were using the interface in English, or reo Māori and English (bilingual use), there was no significant international and national difference in how long the interface pages were being read.

4.3.3 How Many Clients Returned?

Comparisons were made with international return usgMi sessions and national return usgMi sessions to determine if location was a determining factor in whether usgMi clients returned to use the Niupepa digital library. In 2005 there were at least 22 (6.2%) international usgMi clients that undertook more than one session with the Niupepa digital library. There were also at least 60 (4.9%) national usgMi clients that undertook more than one session with the Niupepa digital library. The

international return usgMi clients requested 7,544 (9.7%) pages and documents while national return clients requested 23,743 (15.1%) pages and documents.

Two conclusions can be drawn from these figures. First, the national/international location was not a determining factor in the number of users who returned to use the Niupepa digital library. Second, the national clients who did return to use the Niupepa digital library requested significantly more pages than the international clients who returned. This difference is highlighted when we consider how many requests were submitted per return client. The international usgMi clients requested 33.2 requests per return client. The national usgMi clients requested 59.8 requests per return client. The national usgMi return clients requested almost twice as many pages as the international usgMi return clients.

4.3.4 Which Interface Pages Were Requested?

The percentage of international and national interface pages requested from usgMi, usgEn and usgBi sessions are listed in *Table 4-7*.

When the interface pages requested in the usage sessions (*Table 4-7*) were compared with the interface pages requested in the exploratory sessions (*Table 4-1*) it shows that usage sessions had a higher percentage of requests for the preferences pages. This inclination is understandable as active users were altering the preferences to suit their browsing and searching strategies.

The international usgMi clients were requesting a higher percentage of home pages and preference pages than the national usgMi clients. However, this difference is only 5.1% and 1.3% respectively, a difference that is not statistically significant. The international and national usgEn clients were requesting similar interface pages as the usgMi clients. The usgBi international requests were too few to draw conclusions.

International and National Interface Pages Requested in Usage Sessions						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total pages	407	983	9340	128357	93	458
niupepa home page	84.8%	79.6%	86.7%	80.3%	65.6%	76.9%
NZDL home page	0.0%	1.8%	0.1%	2.2%	0.0%	2.0%
About Eng. Abstract	3.4%	7.0%	4.0%	6.0%	1.1%	5.5%
help page	2.7%	3.8%	2.8%	2.4%	6.5%	4.1%
preferences page	9.1%	7.8%	6.3%	9.0%	26.9%	11.6%
greenstone page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
unknown page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

Table 4-7: International and national comparison of percentages of Niupepa web site interface pages requested in usage sessions.

4.3.5 How Often Was the Language Switched?

Solid conclusions cannot be drawn in comparisons between national and international language switching in usage sessions as there was not enough international language switching data available. Nationally there were 1,512 language switches in 1093 usage sessions; internationally only 206 language switches in 277 usage sessions. Summary international and national switching counts and percentages are displayed in *Table 4-8*.

International and National Switching Summaries						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Number of switches	33	160	187	1057	57	295
reo Māori to English	60.6%	36.9%	90.4%	84.1%	57.9%	55.9%
English to reo Māori	39.4%	63.1%	9.6%	15.9%	42.1%	44.1%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
sessions with switch	6.1%	9.0%	32.5%	20.3%	100.0%	100.0%

Table 4-8: Comparing national and international switching counts and percentages for usgMi, usgEn and usgBi sessions.

It appears that international usgMi sessions had a higher preference for switching from te reo Māori to English (60.6%), whereas national usgMi sessions had a higher preference for switching the interface language from English to te reo Māori (63.1%). International and national usgEn sessions both showed a strong preference to switch from te reo Māori to English (90.4% and 84.1%). International and national usgBi sessions both showed a slight preference to switch from te reo Māori to English (57.9% and 55.9%).

Again, while these assertions can be raised, given the data that has been calculated in *Table 4-8* they cannot be confirmed with any degree of certainty as the count of international switches is too low.

4.3.6 Summary—Interface Behaviours

This section analysed how the Niupepa interface was used by international reo Māori clients and national reo Māori clients. Statistics derived from TLA were used to detail five aspects of how international usgMi clients used the Niupepa interface differently from national usgMi clients.

The statistics revealed that international usgMi sessions averaged 10 minutes, almost half the duration of national usgMi sessions. Although the length in time was less, the usgMi international sessions averaged two more requests per session than the national usgMi sessions. This meant that the international usgMi clients were reading pages more quickly than the national usgMi clients, on average twice as quick. There did not appear to be a significant difference in the number of international and national usgMi clients returning to use the Niupepa collection; however, the national usgMi clients that did return requested almost twice as many pages as the international usgMi clients who undertook return sessions. The international usgMi clients appeared to request a similar proportion of interface pages as the national usgMi clients. While there were too few international usgMi interface language switches (only 33) to draw any firm conclusions, it did appear that the international usgMi sessions had a higher preference for switching the

interface language from te reo Māori to English, whereas national usgMi sessions had a higher preference for switching from English to te reo Māori.

This brief summary completes the comparison of how international and national usgMi clients used the Niupepa interface in 2005. There are two further parts in the comparative analysis of how the Niupepa digital library was used by international and national usgMi clients: a comparative analysis of how documents were accessed and a comparative analysis of how queries were undertaken. The comparative analysis on how documents were accessed by international and national usgMi clients begins in the following section.

4.4 International and National Document Access Behaviours

This section compares how international reo Māori clients and national reo Māori clients accessed documents. A document is defined as a resource within the Niupepa digital library that represents real content, rather than something that provides access to content, such as an interface page. Typically this resource is a newspaper page but it could also be a request for an English Abstract or a request for a Bibliographic Commentary.

Statistics generated from TLA were used to compare four document behaviours of international and national usgMi clients. Document behaviours compared were:

- the methods used to access documents
- the document formats favoured
- the document language favoured
- the apparent read times of documents.

The comparative analysis of these four document behaviours comprises the next four sub-sections.

4.4.1 How Were Documents Accessed?

Three methods were available to clients of the Niupepa digital library accessing documents. The Niupepa documents could be selected from a list of results following a query activity, from a list of publication series and from a list of articles published in a certain date. TLA indicated the methods used to access documents by international and national usgMi clients. The percentages used are displayed *Table 4-9*.

International and National Document Access						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total Documents	5083	17,165	5,955	96,085	224	2,048
From query	84.5%	75.7%	71.5%	82.2%	74.1%	69.8%
From series	10.4%	13.7%	18.0%	5.9%	14.7%	9.8%
From date	2.1%	8.9%	7.1%	9.8%	0.9%	16.8%
Uncertain	3.0%	1.7%	3.5%	2.1%	10.3%	3.7%
p value	2.1*10⁻⁷⁵		7.1*10⁻³⁰⁴		8.1*10⁻¹³	

Table 4-9: *International and national comparison by percentage of how documents were requested by usgMi, usgEn and usgBi clients.*

The query activity was the most popular method of accessing pages, both nationally and internationally for all three usage session types. The very low p values indicate that the differences in how documents were requested by the national and international usage sessions are statistically significant for all types of usage sessions.

The international usgMi clients had a higher preference (84.5%) for using the query facility than the national usgMi clients (75.7%). The national usgMi clients had higher preferences for accessing documents by using the series list and the date facility. The reverse was true for usgEn clients, where national usgEn clients had the highest preference for using a query to access documents.

4.4.2 Which Document Formats Were Favoured?

The newspaper documents stored in the Niupepa collection can be presented to the user in one of three formats, extracted text, full size image, and a preview image. *Table 4-10* lists the numbers and percentages of these three formats with a fourth listing, those requests where it was unclear which page format was requested.

International and National Formats of Requested Documents								
	usgMi		usgEn		usgBi		Total	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total	407	17,165	5,955	96,085	224	2,048	1,262	115,298
Ext. Txt	82.8%	55.5%	85.6%	53.0%	84.4%	51.2%	84.3%	53.4%
Full Size	12.1%	39.5%	8.8%	40.1%	10.3%	38.7%	10.3%	40.0%
Preview	3.5%	4.8%	3.6%	6.4%	3.6%	9.9%	3.6%	6.2%
Undefined	1.6%	0.3%	1.9%	0.5%	1.8%	0.1%	1.8%	0.4%

Table 4-10: International and national comparisons of document types requested by usgMi, usgEn and usgBi clients.

The extracted text format was the most requested format by all client types in both international and national sessions. However, the amount that it was preferred by international and national clients was significantly different. International usgMi clients requested the extracted text format in 82.8% of all document requests, whereas the national usgMi clients requested this format in only 55.5% of documents requests. International usgMi clients exhibited a lesser preference for requesting documents in the full size image format, 12.1% of document requested, as opposed to 39.5% of document requested by national usgMi clients.

The preference differences exhibited by international and national usgMi clients were similar to the preference differences shown by international and national usgEn which, in turn, were similar to those exhibited by international and national usgBi clients.

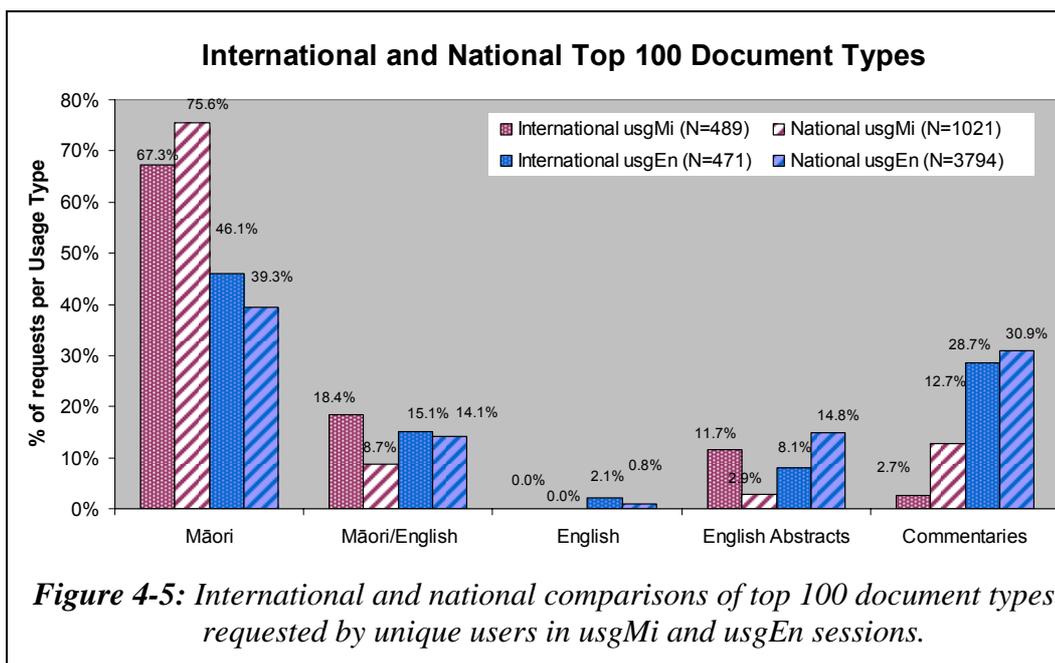
4.4.3 Which Document Types Were Favoured?

Five different types of documents were available to Niupepa clients. These documents consisted of; newspapers written in te reo Māori, newspapers in parallel texts of te reo Māori and English, newspapers written in English, abstracts in English and commentary information written in both English and te reo Māori. A list of document quantity is given in §3.5.3. International and national requests for these document types were compared to determine if international usgMi clients requested different document types from national usgMi clients.

Rather than simply counting the document types requested, the document types requested by unique users were counted. Unique users give a better picture of what a wider range of people accessed and limits the influence in the statistics of the heavy users. It was also decided to collate only the top 100 unique pages requested as limiting the analysis to this number gives a better indication of users' initial preferences, again reducing the influence in the statistics of heavy users of the web site. The data collated in this part of the analysis consisted of:

- 489 unique user requests for 105 pages in international usgMi sessions
- 1,798 unique user requests for 106 pages in national usgMi sessions
- 471 unique user requests for 74 pages in international usgEn sessions
- 3794 unique user requests for 99 pages in national usgEn sessions
- 13 unique user requests for 6 pages in international usgBi sessions
- 233 unique user requests for 91 pages in national usgBi sessions

The usgBi session requests consisted of all the requests for pages by two or more unique users. These counts were considered too low to return valid data so the usgBi sessions were not compared in this particular analysis. The usgMi and usgEn document request counts were calculated as percentages, so that they could be compared, and are graphed in *Figure 4-5*.



Some differences can easily be seen in *Figure 4-5*. International and national usgMi sessions had a stronger preference to request newspaper documents written in te reo Māori. International and national usgEn sessions had a stronger preference to request the commentary information written in English.

The international usgMi clients had stronger preferences than the national usgMi clients to request the parallel reo Māori and English newspaper documents and the English Abstracts. On the other hand, the national usgMi clients had stronger preferences than the international usgMi clients to request the reo Māori newspaper documents and the Bibliographic Commentaries.

The differences noted between the international and national document format preferences of usgMi clients were not apparent to the same extent with international and national usgEn clients.

4.4.4 How Long to Read Documents?

The apparent read times (as explained in §2.5.1) of newspaper documents were compared for international and national usgMi clients. The average apparent read times of full size documents, of extracted text documents and of preview documents were calculated and are displayed for the various usage sessions in *Table 4-11*.

All average reading times in seconds	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
All formats	43.5	79.8	59.2	55.0	57.7	72.5
Full Size Image	77.1	115.1	111.6	93.4	113.0	120.2
Extracted Text	39.3	59.1	55.5	30.3	54.2	46.6
Preview Image	35.9	35.4	37.7	26.3	14.9	24.4

Table 4-11: International and national average apparent read times (in seconds) of documents in usgMi, usgEn and usgBi sessions.

In all types of usage sessions the full size images had the longest average apparent read time. The second longest read averages in all sessions were the extracted text documents. The averages also indicated that the shortest documents to read in all usage sessions were the preview images.

In general the international usgMi clients took less time to read documents than the national usgMi clients. The international usgMi apparent read times were quicker than national usgMi apparent read times for full size images and for extracted text documents and were very similar for preview images.

The reverse was true for usgEn clients. The international usgEn clients averaged longer apparent read times than national usgEn clients for all types of documents.

4.4.5 Summary—Document Access Behaviours

This section compared how international and national reo Māori clients accessed documents. Statistics generated from TLA were used to compare four document access behaviours exhibited by international and national usgMi clients.

The comparisons revealed that using the query facility was the favoured method used to select documents, a method favoured more by international than by national usgMi clients. The extracted text format of documents was preferred by usgMi clients; however, this format was preferred significantly more by international usgMi clients than by national usgMi clients. International usgMi clients displayed higher preferences for requesting parallel reo Māori/English newspaper documents and English Abstracts, whereas national usgMi clients displayed higher preferences for requesting reo Māori newspaper documents and Bibliographic Commentaries. International usgMi clients exhibited quicker apparent read times than national usgMi clients for all types of documents requested.

This brief summary concludes the comparison of how international and national reo Māori clients accessed documents. The comparative analysis of how the Niupepa digital library was used by international and national usgMi clients consists of three parts, a comparative analysis on how the interface was used, a comparative analysis of how documents were accessed and a comparative analysis of how queries were undertaken. The third and final part, how queries were undertaken by international and national usgMi clients, begins in the following section.

4.5 International and National Query Behaviours

This section analyses and compares query behaviours of international and national usage sessions in te reo Māori. A query is defined as the activity undertaken by a client that involves using the search facility of the Niupepa digital library to find a term or string of terms in documents stored in the Niupepa digital library.

Perl scripts have been used in the TLA process to generate query behaviours from the NZDL National and NZDL International web log datasets. The comparative analysis was undertaken on nine query behaviours:

- the number of queries submitted
- the number of query terms used in queries
- the topics queried
- the language of the query terms
- the use of query parameters
- the use of query operators
- the number of documents requested following a query
- the format of documents requested following a query
- the apparent document read times following a query.

These nine query behaviours comprise the next nine sub-sections and detail a comprehensive comparative analysis of how international and national reo Māori clients queried the Niupepa digital library.

4.5.1 How Many Queries Were Submitted?

The numbers of queries submitted by international usgMi clients were compared with the numbers of queries submitted by national usgMi clients. These counts were calculated as percentages and are displayed in *Table 4-12*. Also displayed in *Table 4-12* are the percentages of queries submitted by international and national usgEn clients and by international and national usgBi clients.

International and National Sessions That Contain Queries						
	usgMi		usgEn		usgBi	
Queries/Session	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total Sessions	380	1322	520	4432	14	76
no queries	30.0%	46.4%	46.7%	31.4%	21.4%	19.7%
1–2 queries	35.3%	27.4%	28.8%	32.9%	14.3%	11.8%
3+ queries	34.7%	26.2%	24.4%	35.7%	64.3%	68.4%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table 4-12: International and national comparisons of number of sessions that contain queries in usgMi, usgEn and usgBi sessions.

International usgMi clients were more likely to submit queries than national usgMi clients. International usgMi clients submitted queries in 70.0% of their sessions, while national usgMi clients submitted queries in only 53.6% of their sessions. When usgMi sessions did involve a query activity, half of these sessions comprised just one or two queries and the remainder comprised three or more queries.

The likelihood of an international and national client submitting a query was reversed when the interface language was set to English. International usgEn clients submitted queries in only 53.2% of their sessions, while national usgEn clients submitted queries in 68.6% of their sessions. The international usgBi sessions appeared to submit queries in a similar proportion to the national usgBi sessions; however, because of the low international usgBi session count (there were only 14) this conclusion remains indeterminate.

4.5.2 How Many Query Terms Were Submitted?

A count was taken of the number of query terms submitted with each international and national usgMi query. The percentages, the number of queries, and the chi square p values are displayed in *Table 4-13*.

International and National Number of Query Terms per Query						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total Queries	1399	3368	1324	15,907	82	422
1 term/query	42.5%	48.6%	58.3%	48.9%	50.0%	46.4%
2 terms/query	30.3%	29.3%	24.8%	32.4%	36.6%	32.2%
3+ terms/query	27.2%	22.0%	16.8%	18.7%	13.4%	21.3%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
p value	$6.7 * 10^{-5}$		$1.2 * 10^{-10}$		0.26	

Table 4-13: International and national percentage comparisons for the number of terms in a query for usgMi, usgEn and usgBi sessions.

A comparison of the national and international percentages in the usgMi sessions seems to suggest that the proportions are similar. However, with a chi square p value of $6.7 * 10^{-5}$ there is a significant difference in the number of query terms submitted in international and national usgMi queries. International usgMi queries contained a higher proportion of queries with three or more query terms. National usgMi queries contain a higher ratio of queries with just one query term.

The reverse was true for usgEn sessions. International usgEn queries contained a higher proportion of queries with just one query term, while national usgEn queries contained a higher proportion of queries with three or more query terms.

4.5.3 What Topics Were Queried?

The topic categories of the query terms were analysed to see if there was a difference between query terms submitted internationally and those submitted nationally. The topic categories were not analysed to the language of session level as the counts of terms became too low to draw any meaningful conclusions. Thus, international and national comparisons were made but international usgMi/national usgMi comparisons were not analysed.

International and National Query terms		
	International Terms	National Terms
1	haka (28)	whakatauki (68)
2	harakeke (21)	makaka (53)
3	waiata (15)	Whakaki (48)
4	koingo (12)	awaitaia (41)
5	tutakangahau (12)	matariki (36)
6	aroha (10)	tioriori (35)
7	mangamaunu (10)	waiata (35)
8	hawaii (9)	parihaka (31)
9	hetet (9)	Parihaka (28)
10	churton (8)	Te wharepu (28)
11	Hetita (8)	treaty (27)
12	R.T.K. (8)	oruanui (26)
13	India (8)	Tapiri (24)
14	Immigration Lists (7)	motakotako (24)
15	pounamu (7)	tawhiao (24)
16	urewera (7)	treaty of waitangi (23)
17	davis (7)	Ratana (23)
18	Immigration Ships from Britain (7)	kingitanga (23)
19	tuhoe (7)	arawa (22)
20	Newspapers (7)	"friend (22) ⁱ
21	india (7)	utiku potaka (22)
22	whiro tirea (6)	pipiwharauoa (20)
23	dog tax (6)	rangatiratanga (19)
24	whatu (6)	Tuwhenua (19)
25	tauanui (6)	Chinese (19)
26	taniwha (6)	ngapuhi whakatauki (19)
27	"tukua te rangi" (6)	

Table 4-14: *The query terms (with counts) most commonly submitted by international and national clients.*

ⁱ This search term included an opening speech mark (but not a closing one).

Internationally, 2,803 queries were submitted which consisted of 2,038 different query strings. Nationally, 19,668 queries were submitted which consisted of 12,204 different query strings. Query terms that fell into the top 25 most often searched terms are displayed in *Table 4-14*, with the number of times that the term was requested written in brackets after the term.

Of the query terms listed above, only one term, waiata (*song*), is listed in both the national and international top 25 query terms.

As mentioned in §3.6.3, the topics of the search terms were categorised so that they could be compared to determine international and national differences. The top 100 unique query terms were examined and categorised into the topic headings. For international usage sessions this involved 132 terms that were queried 614 times. For national usage sessions this involved 103 terms that were queried 1761 times. These terms were categorised into the topic categories and then the percentages of terms per category were calculated. These query topic percentages are displayed in *Figure 4-6*.

Figure 4-6 shows that the topic categories searched for by international and national clients were similar. The topic most searched for was names of people followed by names of places. The national clients showed significantly higher preferences for searching about tribes and nationalities, and publications.

In the top 100 unique query terms, only five terms were common to searches undertaken by both international and national clients. Those common query terms were, in order of frequency, waiata [*song*] (15 international requests, 35 national requests), haka [*dance with postures*] (28 international requests, 15 national requests), tuhoe [*a tribe located in the Central North Island of Aotearoa*] (15 international requests, 7 national requests), māori (4 international requests, 12 national requests), and kingitanga [*the kingship*] (3 international requests, 10 national requests).

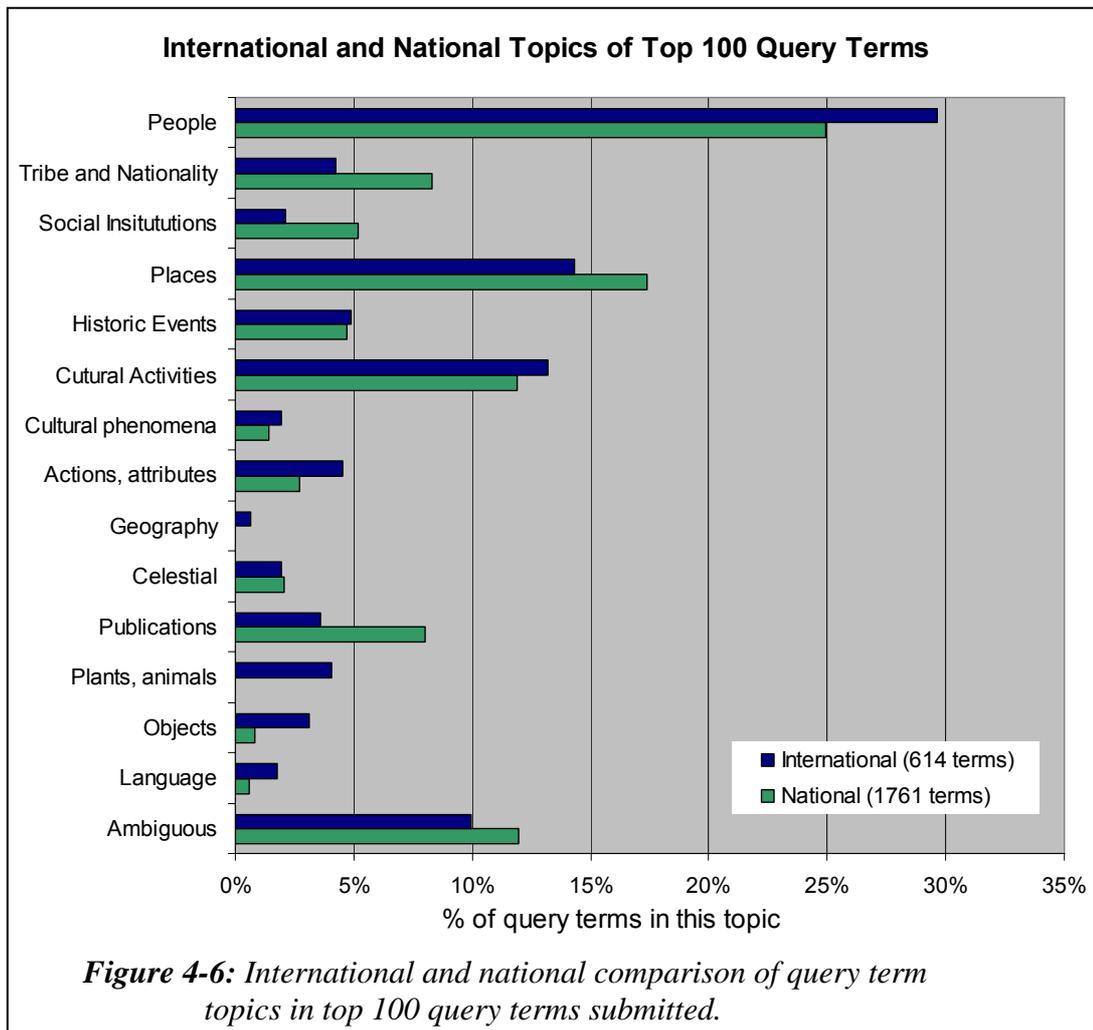


Figure 4-6 suggests the categories searched by international and national users are similar. A comparison of the top 100 query terms shows that there was a 95% difference in the terms used by international and national clients. Different query terms can be expected in the people and places categories as clients will search for names and places that are individual to them. However, similar query terms could be expected for the other categories. The fact that these query terms are not the same suggests that international and national clients were using different vocabularies to retrieve similar topics of information.

4.5.4 What Was the Language Of the Query Terms?

The international and national datasets were examined to determine what language was used in the query terms submitted to the Niupepa digital library in 2005. In total there were 14,160 disparate query terms submitted in 22,471 queries. Most query terms were written in te reo Māori, with almost all of the remaining terms written in English. Seven query terms consisted of numbers, presumably dates and, for the purpose of this analysis, these terms were not considered as reo Māori or English terms.

The language of the most common query terms submitted by international usgMi clients was compared with the language of the most common query terms submitted by national usgMi clients. There were 167 international usgMi query terms examined, submitted in 437 queries, with 94 national usgMi query terms examined, submitted in 532 queries.

The international usgMi clients submitted 95.9% of their query terms in te reo Māori. The national usgMi clients submitted 90.0% of their query terms in te reo Māori. These figures show that the international usgMi clients had a higher (5.9%) preference for submitting query terms in te reo Māori.

4.5.5 Were Query Parameters Altered?

Clients who undertook queries had the facility to alter some query parameters on the Preferences page and on the Query page. On the Preference page, clients were able to alter parameters such as query mode, query style, case matching, word endings and search history (see *Figure 3-2*). On the Query page, clients were able to alter whether the results of the search must contain *some* or *all* of the query (see *Figure 3-3*).

The query parameters located on the Preference page were not altered by international or national usgMi clients. Only 0.9% of international usgMi clients and 0.2% of national usgMi clients altered one of these parameters.

The Some/All query parameter located on the Query page, a page used by clients to submit queries, was altered by international and national usgMi users. There were 11.0% of international usgMi clients and 7.6% of national usgMi clients who altered the Some/All query parameter. These percentages indicate that international usgMi clients had a higher preference to alter this query parameter than the national usgMi clients.

In contrast, the reverse preference was noted in international and national usgEn clients. There were 5.9% of international usgEn clients and 10.9% of national usgEn clients who altered the Some/All query parameter. These percentages indicate that national usgEn clients had a higher preference to alter this query parameter than the international usgEn clients.

4.5.6 Were Query Operators Used?

Two types of query operators were available to clients of the Niupepa digital library in 2005; logical query operators (and, or, not, *, +, -) and phrase defining operators (quotation marks). Less than 2% of international and national queries contained the use of logical operators; consequently, the use of these operators was not investigated further. However, it was a different matter for phrase defining operators.

There were 344 queries (out of a total 1,399 queries) submitted by international usgMi clients that included the use of quotation marks to specify phrase searching. This figure represents one in four (24.6%) of all the queries submitted by international usgMi clients and is a considerable proportion. There were 481 queries (out of a total 3,368 queries) submitted by national usgMi clients that included the use of quotation marks. This figure represents one in seven (14.3%) of all national usgMi queries.

Queries submitted by international usgEn clients also showed a higher presence of phrase defining operators (8.9%) than queries submitted by national usgEn clients

(6.2%). The difference observed in international and national usgEn clients was not as large as the difference observed in international and national usgMi clients.

4.5.7 Documents Requested Following a Query

The numbers of documents requested following the submission of a query were examined for international and national clients who used the interface in te reo Māori. There were 1,399 queries submitted by international usgMi clients that led to the request of 4,213 documents; thus, 3.0 documents were requested for every international usgMi query. There were 3,368 queries submitted by national usgMi clients that led to the request of 11,868 documents; thus, 3.5 documents were requested for every national usgMi query. The national usgMi clients requested more documents following a query than the international usgEn clients.

As a comparison, there were 3.2 documents requested following an international usgEn query and 4.7 documents requested following a national usgEn query. These figures indicate that the national usgEn clients requested significantly more documents following a query than the international usgEn clients.

4.5.8 Formats Favoured Following a Query

There were 8,572 international and 88,290 national requests for documents following a query. The percentages of document formats requested by international and national usgMi clients following a query are displayed in *Table 4-15*. Also listed are the percentages of document formats requested by international and national usgEn and usgBi clients.

International and National Document Formats Requested After a Query						
	usgMi		usgEn		usgBi	
	Intern.	Nat.	Intern.	Nat.	Intern.	Nat.
Total	4,213	11,868	4193	74,999	166	1,423
Extracted Text	89.0%	62.9%	90.2%	55.0%	90.4%	51.7%
Full Size Image	8.8%	33.0%	6.8%	40.1%	6.6%	36.5%
Preview Image	1.9%	4.0%	2.2%	4.7%	2.4%	11.6%

Table 4-15: International and national percentages of document formats requested following a query.

International usgMi clients requested a higher proportion (89.0%) of extracted text documents following a query than national usgMi clients (62.9%). Alternatively, national usgMi clients requested a higher proportion of full size images (33.0%) and preview images (4.0%) than international usgMi clients (8.8% and 1.9% respectively).

The preference in the international usage sessions to request the higher proportion of extracted text documents and the national usage sessions to request the higher proportion of images following a query was also evident in the usgEn and usgBi sessions.

4.5.9 Document Read Times Following a Query

The apparent read times of documents requested following a query were examined for national and international usgMi sessions. This comparison is similar to the comparison undertaken in §4.4.4; however, in this instance only the documents requested following a query were used to calculate the apparent read times. The average apparent read times for each type of document requested are displayed in *Table 4-16*.

International and National Apparent Read Times after a Query

Average Apparent Read (Seconds)	International usgMi		National usgMi	
	Non Query	After Query	Non Query	After Query
All Documents	58.5	40.5	92.1	74.7
Full Size Image	78.7	76.2	130.6	104.2
Extracted Text	58.1	37.1	49.9	61.5
Preview Image	32.6	43.9	24.4	43.6

Table 4-16: Apparent read times (seconds) by international and national usgMi clients following the submission of a query.

Table 4-16 shows that in general usgMi clients, both nationally and internationally, took less time to read documents following the submission of a query. The decrease in apparent read times of all documents by international and national usgMi clients was approximately 18 seconds.

While both international and national clients displayed a similar decrease in apparent read times for all documents, when the document types were analysed individually there were distinct differences. International usgMi clients took about the same time to read full size images, less time to read extracted text documents and more time to read preview images following a query. National usgMi clients took less time to read full size images, more time to read extracted text documents and more time to read preview images following a query.

4.5.10 Summary—Query Behaviours

This section analysed and compared query behaviours of international and national usgMi clients. Statistics produced by the TLA process were used to examine nine query behaviours exhibited by international and national usgMi clients.

These statistics revealed that international usgMi clients were more likely to submit queries than national usgMi clients. When queries were submitted,

international usgMi clients were more likely to search for multiple query terms. International and national clients searched for similar topics but they also used a different vocabulary of query terms. The international usgMi clients entered a higher percentage of query terms in te reo Māori, were more likely to alter the Some/All query parameter and used the phrase defining operator (the quotation marks) almost twice as often as the national usgMi clients. Both of these behaviours are perhaps a consequence of usgMi clients submitting a higher number of terms per query. Fewer documents were requested by international usgMi clients following a query but they requested a higher proportion of extracted text documents. Generally, both client types spent less time reading documents following a query, except in the case of preview images which showed longer reading times following a query.

A comparative analysis of how international and national usgMi clients *used* the Niupepa digital library has been discussed by examining three behaviours, including query behaviours. Before this chapter is complete, one further comparison merits discussion; how international and national usage differences in te reo Māori compared with international and national usage differences in English. This question is addressed in the following section.

4.6 Te Reo Māori Usage/English Language Usage

The purpose of this chapter has been to determine if, and subsequently define how, te reo Māori was used differently by international and national clients of the Niupepa digital library. Given that international and national usgMi differences were noted, another question springs to mind. Were similar international and national differences displayed by clients who used the Niupepa digital library in English? If the international and national differences occurred in a similar manner in English perhaps these differences are likely to occur for all languages? If this is the case then strategies undertaken to increase the international participation in one language can be applied to other languages.

The comparison of international and national usgMi and usgEn behaviours has been condensed into three sections, a comparison of interface behaviours, a comparison of document access behaviours and a comparison of query behaviours.

4.6.1 Interface Behaviour Differences

International usgEn sessions were shorter than national usgEn sessions in both time and number of requests. The apparent read times of interface pages by international and national usgEn clients were similar. National usgEn clients showed interest in the NZDL web site home page and the Niupepa preferences page whereas the international usgEn clients did not show the same interest in these pages.

When international and national usgEn clients switched the language of the interface, both had a high preference to switch the language to English.

The international and national usgEn interface behaviour comparisons were different from the international and national usgMi interface behaviour comparisons in almost all aspects. The only similar aspect was that international usgMi sessions were shorter in time than national usgMi sessions; all other comparisons showed dissimilar results. The international and national usgMi read times were different, the international and national usgMi clients requested proportionally the same interface pages, and when the international and national usgMi clients switched languages they did so in different ways.

4.6.2 Document Access Behaviour Differences

International usgEn clients had a lower preference than the national usgEn clients for using the query facility to access documents and a subsequently higher preference for selecting documents by browsing the publication series list.

Extracted text documents were the formats preferred more by international usgEn clients than by national usgEn clients, with national usgEn clients displaying a higher preference for viewing full size images. There was little difference in the types of documents preferred by international and national usgEn clients. The international usgEn clients averaged longer apparent read times for all types of documents.

The international and national usgEn document access behaviours were different from the international and national usgMi document access behaviours in three out of four aspects. Preferences for the type of document were the only similarity.

4.6.3 Query Behaviour Differences

International usgEn clients were less likely to submit queries than national usgEn clients. International usgEn queries contained a higher proportion of queries with just one query term, while national usgEn queries contained a higher proportion of queries with two or more query terms. There were 5.9% of international usgEn clients who altered the Some/All query parameter compared with 10.9% of national usgEn clients. Queries submitted by international usgEn clients showed a higher presence of phrase defining operators (8.9%) than queries submitted by national usgEn clients (6.2%). International usgEn clients requested significantly fewer documents following a request than the national usgEn clients (3.2 compared to 4.7). International usgEn clients requested a higher proportion of extracted text documents following a query and a lower proportion of full size and preview.

The international and national usgEn query comparisons were different from the international and national usgMi query comparisons in almost all respects. Of the six international and national usgEn query comparisons discussed, only one, the types of documents requested following a query, was similar to the query behaviours noted in the international and national usgMi comparative analysis.

4.6.4 Summary—Behaviour Differences

The results of analysing the international and national usgEn behaviours show there were distinct differences in behaviours exhibited by international and national usgEn clients. These differences occurred in interface behaviours, document access behaviours and query behaviours. A conclusion that can be drawn is that the information retrieval strategies used by international clients who used the Niupepa digital library in English were different from those used by national clients who used the Niupepa digital library in English.

The differences displayed by international and national usgEn clients were not the same as those displayed by international and national usgMi clients. Thirteen of sixteen international and national usgEn behaviours were either the complete opposite of or significantly different from the international and national behaviours exhibited by usgMi clients. These results suggest that usage of the Niupepa digital library by international clients in English is quite different from usage by international clients in te reo Māori. Consequently, strategies undertaken to increase the international participation in one language cannot be applied to another language.

4.7 Summary

This chapter began by asking if te reo Māori was used by international clients of the Niupepa collection. §4.1 showed that international clients were responsible for only 10% of the requests and were more likely to explore but then not go on to use the Niupepa digital library. However, when international clients did use the Niupepa digital library they were three times more likely to use the digital library in te reo Māori than were national clients.

Exploratory activities of international expMi clients were compared with exploratory activities of national expMi clients in §4.2. International expMi sessions were shown to be 20% longer (in time) with 27% longer apparent read times than the national expMi sessions. International expMi clients did not show the same interest in other pages of the Niupepa interface as did their national counterparts.

Usage activities of international usgMi clients were compared with usage activities of national usgMi clients in three areas; interface behaviours (§4.3), document access behaviours (§4.4) and query behaviours (§4.5). International usgMi sessions tended to be quicker, with more requests but comparatively fewer interface language switches than national usgMi sessions. International usgMi clients were more likely to use queries but requested fewer documents following a query. International usgMi clients requested more extracted text formats and parallel (reo Māori/English) and English language documents than national usgMi clients. The higher proportion of queries undertaken by international usgMi clients included a higher proportion of multiple query terms, more use of the Some/All parameter and the phrase defining operator. These differences were all statistically significant and revealed that te reo Māori was not used in the same way by international clients of the Niupepa digital library as by national clients.

A further question was considered in the preceding section (§4.6): given that te reo Māori was used in different ways by international and national clients, was the English language also used differently? Results from the international and

national usgEn comparison were completely opposite to or significantly different from results from the international and national usgMi comparison in 13 of the 16 behavioural differences analysed. Consequently, international and national differences that occurred in reo Māori usage were not the same as international and national differences in English language usage.

4.7.1 Discussion

The Niupepa web site is likely to hold more interest for national clients, users based in Aotearoa, because of the nature of the content. Historical newspaper documents are more likely to be of interest to people from the country in which the newspapers were published. Further, if the newspapers are mostly written in an indigenous language, then they are of more interest to the country where that indigenous language is spoken. This higher interest has been evident in the statistics generated when comparing national and international requests. Almost 85% of all filtered requests and 90% of all usage requests to the Niupepa web site were requests from (national) clients in Aotearoa.

It was difficult to determine exactly where the international usgMi clients were from. Statistics discussed in §3.4.4 indicated that most international requests submitted by the top 100 IP Addresses were coming from Australia and the United States of America. However, only 2.7% of requests from Australia and 5.9% of requests from USA were catalogued as usgMi requests. A large number of international usgMi requests did appear to originate from the islands of Hawai'i but as only the top 100 IP Addresses were identified it is difficult to confirm if this number was significant.

The international clients who used the Niupepa digital library in te reo Māori have been shown to be quite a different user group than the national clients who used the Niupepa digital library in te reo Māori. The increase in international usgMi sessions when the default language was set to te reo Māori (see *Figure 4-3*) and the relatively low number of interface switches indicated that international usgMi

clients were bilingual users who were appeared confident in using the interface in te reo Māori. The international usgMi clients appeared to be experienced in information retrieval; their preference for queries and their use of query parameters, query operators and multiple query terms indicated they were familiar with query activities. The fewer documents requested following a query suggested that their query activity was more focused and/or more accurate. Their short usage session times but high usage session requests indicated they were precise in obtaining the information that they were seeking. The higher proportion of requests for parallel (reo Māori/English) and English language documents suggested that international usgMi clients were seeking information for English language purposes more often than were national usgMi clients.

The international usgMi clients of the Niupepa digital library appeared to be experienced information retrieval clients who were confident to use te reo Māori. These clients were able to take advantage of electronic reo Māori resources made available on the Internet. Use of te reo Māori by international clients in the Niupepa digital library is an example of how a resource such as the Niupepa digital library can be used to assist in the international usage of an indigenous language.

5 The Effect of the Default Language Setting On Te Reo Māori Usage

This chapter examines how the default language setting of the Niupepa digital library affected use of te reo Māori by Niupepa clients. Did setting the default language to reo Māori (defMi) increase use of the digital library in te reo Māori, as proposed in preliminary research (Keegan & Cunningham, 2005)? Did setting the default language to reo Māori change the usage behaviours of te reo Māori clients who used the Niupepa digital library? In this chapter statistics derived from TLA will be used to seek answers to these questions.

The default language is defined as the language that the Niupepa web site was displayed in when the home page (www.nzdl.org/niupepa) was requested. Default language analysis is possible because throughout 2005 the default language setting of the Niupepa web site was alternated, on a monthly basis, between te reo Māori and English, as described in §2.2.3.

Statistics generated from Transaction Log Analysis (TLA) are used to compare te reo Māori usage for the default language settings of reo Māori and English. These statistics show the information retrieval behaviours of te reo Māori clients were significantly different for the different default language settings. These differences are discussed in sections about the amount of reo Māori use, exploratory behaviours, usage of the Niupepa interface, accessing documents and undertaking queries.

5.1 Did Te Reo Māori Usage Increase?

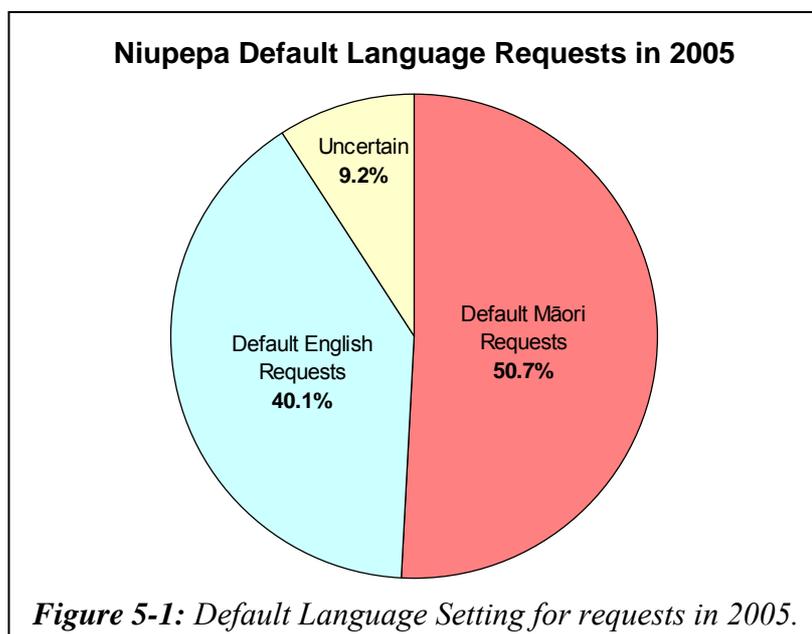
When the default language of a bilingual interface is set to an indigenous language, it can be expected that use of the interface in the indigenous language will increase. Bilingual clients are likely not to change the language of the interface and monolingual clients who do not understand the indigenous language will register at least one request in the indigenous language before the interface is changed to the familiar (non-indigenous) language. But exactly how many more indigenous language requests were there when the default language was set to the indigenous language? Did this setting discourage other (non-indigenous) clients from using the digital library? How does this setting affect the chronological behaviours of clients accessing the digital library? These questions will be addressed in the next five sub-sections.

5.1.1 Effect on the Number of Niupepa Requests

The Niupepa web site received 185,001 filtered requests in 2005. Of these:

- 93,867 were received when reo Māori was the default language (defMi)
- 74,186 were received when English was the default language (defEn)
- 16,948 were received when it was uncertain which interface language was set as the default.

The numbers of requests recorded with the different default language settings have been calculated as percentages and are displayed in *Figure 5-1*.



Interestingly, there were significantly fewer web robot requests when the default interface language was set to te reo Māori. There were 28.8 web robot requests per day with a defMi setting and 75.8 per day with a defEn setting. This difference reveals that web robots submit almost three times more requests to a reo Māori/English bilingual web site when the default language of the web site is set to English.

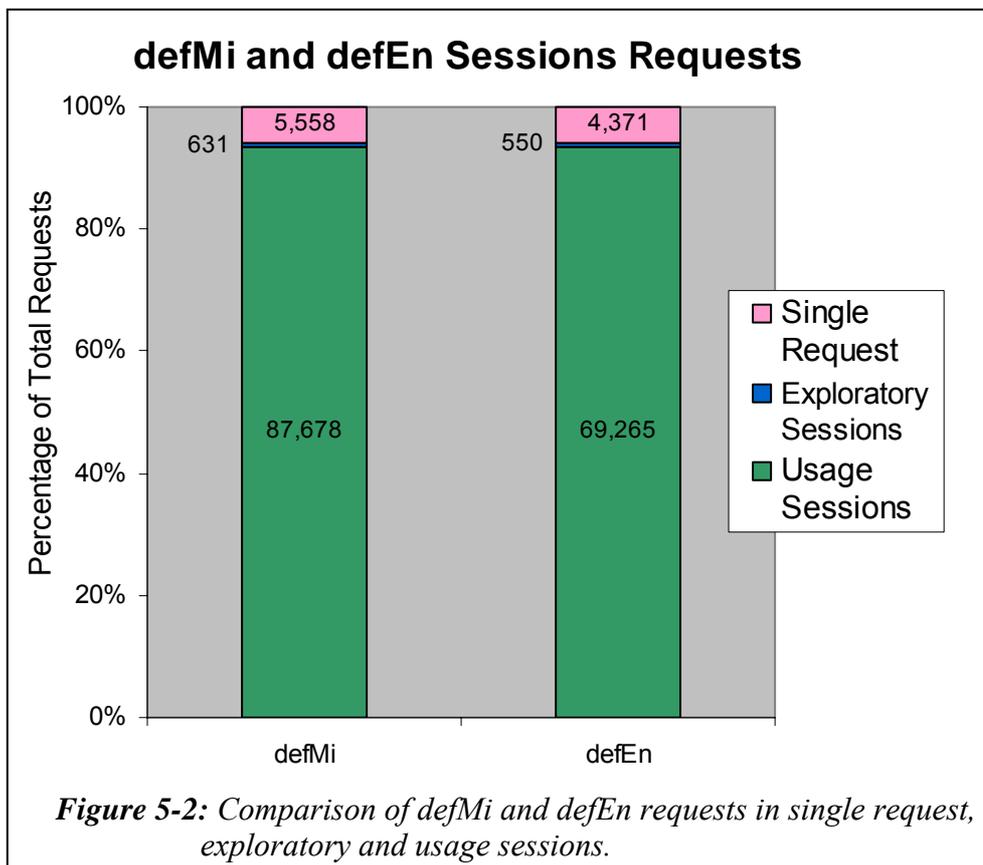
The 93,867 defMi requests were recorded in 198 days at a rate of 474.1 requests per day. The 74,186 defEn requests were recorded in 137 days at a rate 541.5 requests per day. Setting the Niupepa digital library default language to te reo Māori resulted in 12.5% fewer requests overall.

5.1.2 Were Potential Niupepa Clients Discouraged?

Requests recorded when the default language was set to te reo Māori (defMi) and when the default language was set to English (defEn) were classified into single request sessions, exploratory sessions and usage sessions. If the default language

setting significantly affected the number of potential clients then this would be apparent in the proportion of single request, exploratory and usage sessions. A higher proportion of defMi single request and exploratory sessions coupled with a lower proportion of defMi usage sessions would indicate fewer potential clients were using the Niupepa collection because of its default language being set to te reo Māori.

The numbers of requests in defMi and in defEn sessions are displayed together in *Figure 5-2*.



The defMi and defEn showed a remarkable similarity in the proportion of requests per sessions type. There were 5,558 defMi requests (5.92%) and 4371 defEn requests (5.89%) classified as single requests. There were 631 defMi requests (0.67%) classified into 277 defMi exploratory sessions and 550 defEn requests (0.74%) classified into 237 defEn exploratory sessions. There were 87,678 defMi

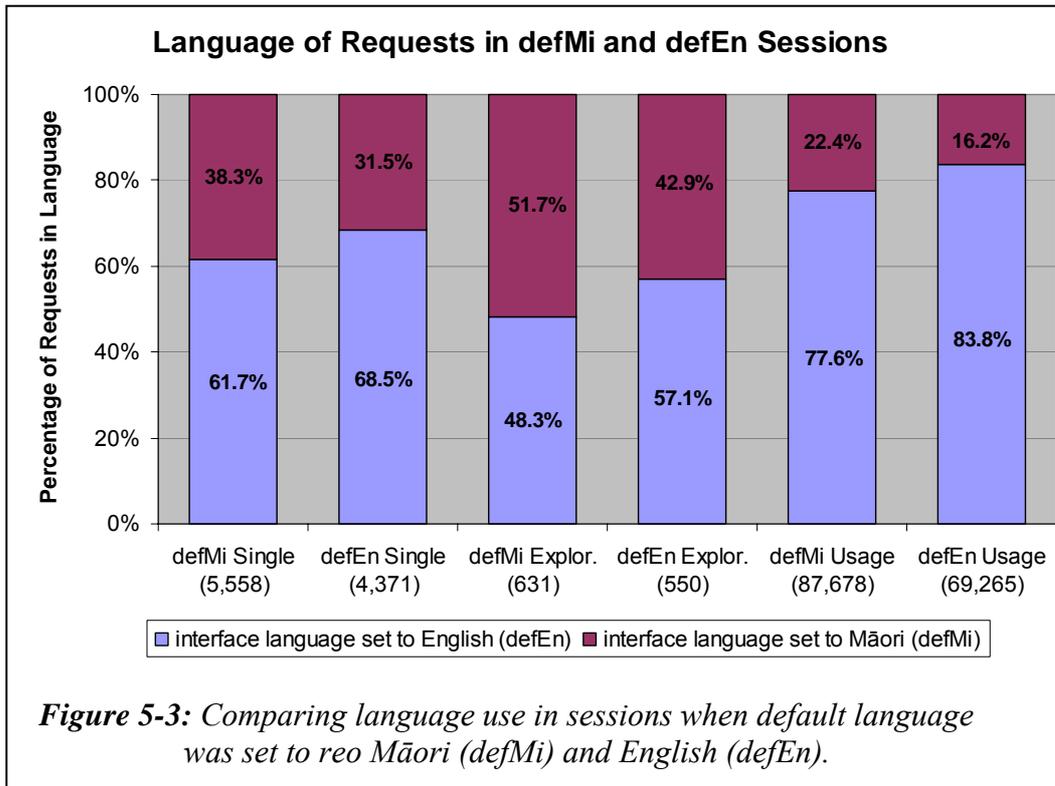
requests (93.41%) classified into 3,238 defMi usage sessions and 69,265 defEn requests (93.37%) classified into 2,971 defEn usage sessions. Chi square analysis of request counts gives a p value of 0.24, indicating no significant difference in the proportions of defMi and defEn session counts.

Setting the default language of the Niupepa digital library to reo Māori did not appear to discourage potential Niupepa clients. There were the same proportion of exploratory/usage sessions by Niupepa clients when the default language was set to reo Māori as there were when the default language was set to English.

5.1.3 Effect on Number of Te Reo Māori Requests

The number of requests submitted in te reo Māori when the default language was set to reo Māori was compared with the number of requests submitted in te reo Māori when the default language was set to English. These numbers have been calculated as percentages and are displayed in *Figure 5-3*.

For the three session types, the proportion of requests in te reo Māori was higher when the default language was set to reo Māori. For single request sessions there was a 6.8% higher proportion of requests in te reo Māori with a defMi setting (38.3% compared with 31.5%). For exploratory sessions there was a 8.8% higher proportion of requests in te reo Māori with a defMi setting (51.7% compared with 42.9%). For usage sessions there were a 6.2% higher proportion of requests in te reo Māori with a defMi setting (22.4% compared with 16.2%). Overall the percentage of requests in te reo Māori increased by 6.2% when the default language of the interface was set to te reo Māori.



As well as the percentage of requests in te reo Māori, the number of requests in te reo Māori was also examined. The requests show an average of 99.1 reo Māori usage requests per day with a defMi setting and 81.9 reo Māori usage requests per day with a defEn setting. There was also an average of 423.7 English usage requests per day with a defEn setting and 343.7 English requests per day with a defMi setting.

Setting the Niupepa digital library default language to te reo Māori resulted in a 21% increase in Niupepa requests in te reo Māori.

5.1.4 Effect on Number of Te Reo Māori Clients

The statistics produced by the TLA process showed counts of how many unique identifiers had been generated for the different session types (§2.4.1 has a discussion on unique identifiers). Unique identifier numbers cannot give an accurate figure of unique client numbers because of behaviours such as disabling

cookies and multiple users sharing single accounts. However, unique identifiers indicate client numbers, which can then be used for comparison purposes.

In the 137 days that the default language of the interface was set to English, 580 unique identifiers were recorded in reo Māori usage sessions. In the 198 days that the default language of the interface was set to reo Māori, 883 unique identifiers were recorded in reo Māori usage sessions. This equates to 4.2 unique identifiers per day with a defEn setting and 4.5 unique identifiers per day with a defMi setting. Setting the default language of the interface to reo Māori did appear to increase the numbers of unique clients in te reo Māori but not to a significant extent.

In contrast, the default language setting did appear to affect the number of usage clients in English by a significant amount. There were 15.8 unique identifiers per day in English usage sessions with a defEn setting but 10.1 unique identifiers per day in English usage sessions with a defMi setting. Setting the default language of the interface to reo Māori appeared to decrease by 36% the numbers of unique clients who used the Niupepa digital library in English.

5.1.5 Effect on When Usage Occurred?

Requests recorded in defMi and defEn usage sessions have been plotted chronologically and are shown in *Figures 5-4, 5-5 and 5-6*. The sessions have been plotted as percentages of requests per session to take into account the variation in frequency of requests between the two default language settings.

Figure 5-4 shows the requests as they occurred in a 24 hour time period. The requests occurred in similar time periods no matter which default language was set. Most of the requests were received during regular 9:00am to 5:00pm work hours, with a small peak again in the evening, probably after dinner time.

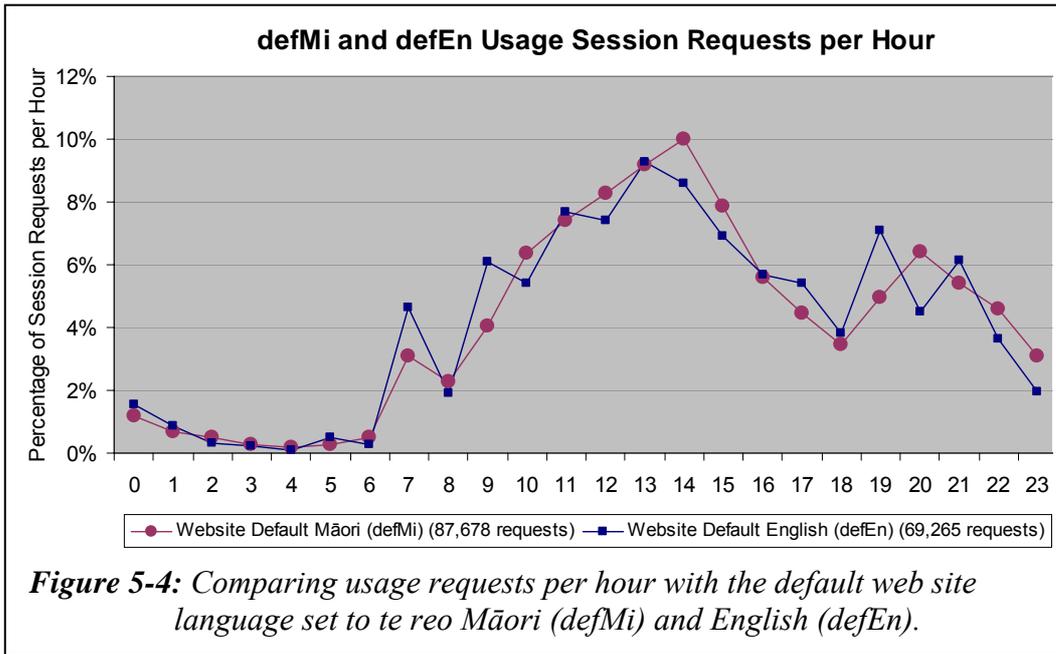
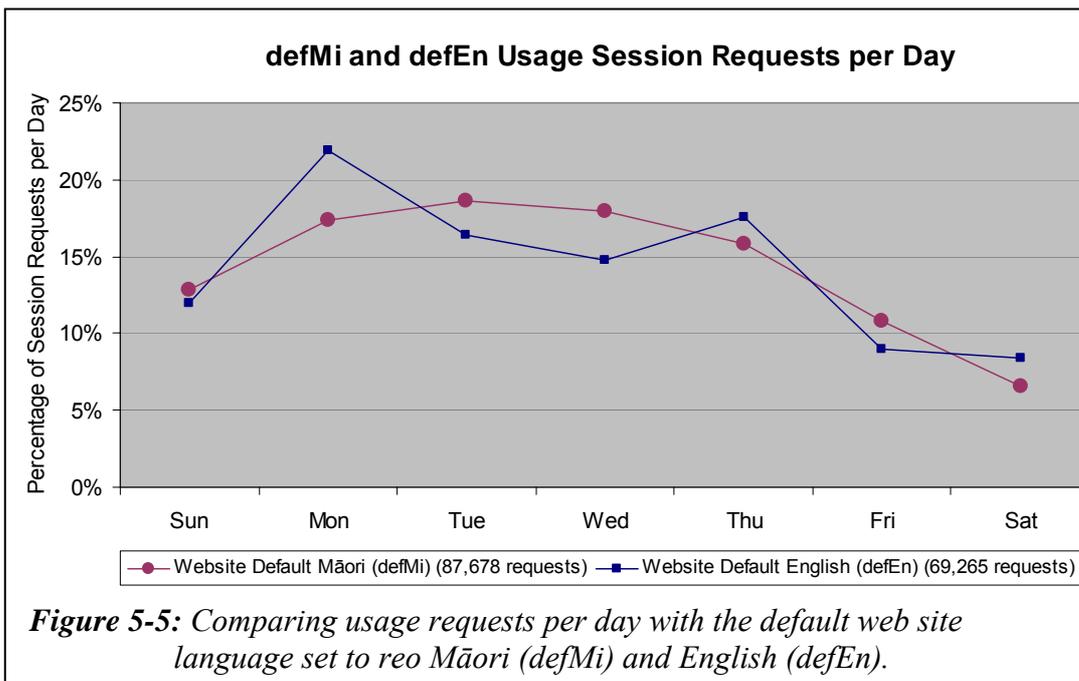


Figure 5-5 shows the requests as they occurred each day. Again, it can be seen that the requests occurred in similar time periods irrespective of the default language setting. Most of the activity occurred from Monday to Thursday, with the least activity on a Saturday.



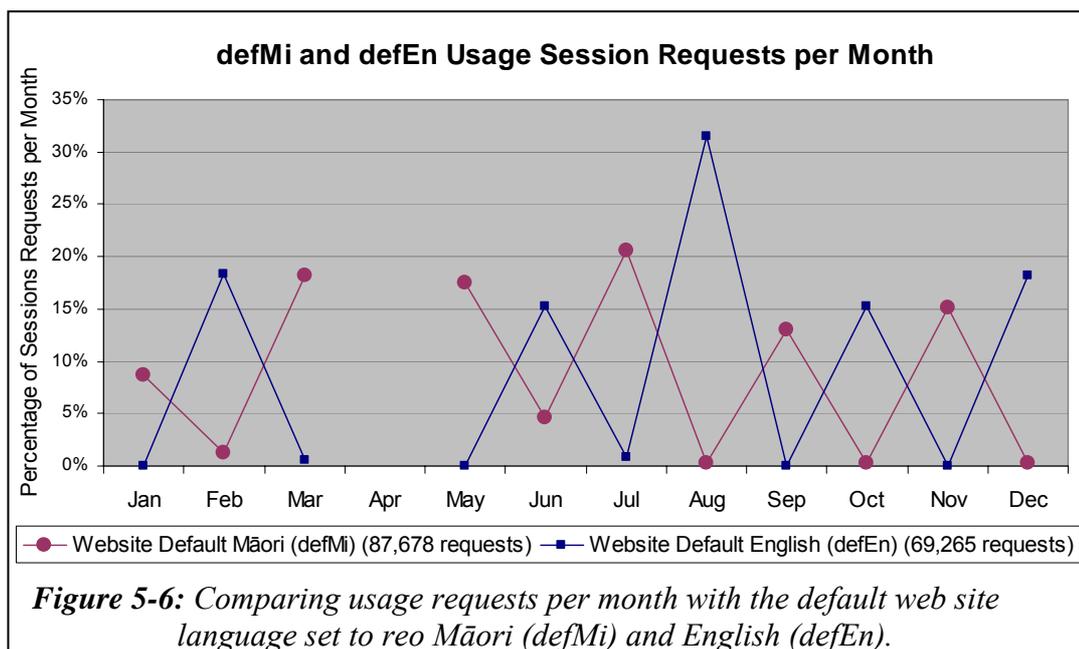


Figure 5-6 shows the requests as they occurred each month. The most requests in te reo Māori corresponded with the months that the default language was defMi. Similarly, months with the defEn setting corresponded with the most requests in English. April shows no requests as they were eliminated from the dataset because of the uncertainty of the default language setting (see §2.2.3). June shows requests from both datasets, indicating that the switching of the default language was late in this month.

The default language setting of the web site did not affect the hours and days of the week that requests were received. Requests recorded each month accurately reflected when the default language setting was alternated.

5.1.6 Summary—Effect on Te Reo Māori Usage

This section has used statistics from TLA to compare the number of reo Māori requests received during a defMi setting with the number received during a defEn setting. The statistics have shown that there was a 6.2% higher reo Māori/English language ratio and 21% more reo Māori usage requests when the default language

was set to te reo Māori. These requests were made by approximately the same number of clients, indicating that the reo Māori default language setting did not encourage new clients to use te reo Māori; rather it encouraged bilingual clients to use te reo Māori more often.

While there was an increase in reo Māori requests there was also a decrease in the number of requests overall. When te reo Māori was used as the Niupepa digital library's default language setting there was a 12.5% reduction in the overall number of requests to the digital library. The proportion of requests in single request, exploratory and usage sessions was the same for both default language settings. This similar proportion implies that the default language setting did not influence the number of exploratory clients who then did not go on to use the Niupepa digital library. Exploratory behaviours of clients are examined in more detail in the next section.

The statistics have also shown that the default language setting of the Niupepa digital library did not appear to influence when the digital library was used by Niupepa clients.

5.2 Effect on Exploratory Behaviours

This section examines how the default language setting affected clients who explored the Niupepa digital library but then did not go on to use it. Although exploratory sessions represented just less than 1% of the total requests, they did give insights into how setting the default language to reo Māori (defMi) or English (defEn) affected the behaviour of exploratory clients.

A high proportion of exploratory sessions were less than two minutes in duration; 80.9% of defMi sessions and 78.1% of defEn sessions. A high proportion of exploratory sessions consisted of just two requests; 83.0% of defMi and 79.7% of defEn sessions, many of which simply involved a request for the home page in one language, then a request for the same page in the other language. With p

values of 0.67 and 0.92 there was no significant statistical difference in the length of defMi and defEn exploratory sessions, either in duration or number of requests.

The defMi and defEn exploratory sessions were then further subdivided by language category. There were:

- 232 requests in 99 defMi expMi sessions at 1.2 requests per day
- 167 requests in 69 defEn expMi sessions at 1.2 requests per day
- 399 requests in 178 defMi expEn sessions at 2.0 requests per day
- 383 requests in 168 defEn expEn sessions at 2.8 requests per day

These figures show that the default language did not have any effect on the number of clients who used te reo Māori to explore (but then not use) the Niupepa digital library. However, it did appear to affect the number of clients who used English to explore the Niupepa digital library; there appeared to be 28.6% fewer expEn requests when the default language of the interface was set to te reo Māori.

5.2.1 Effect on Interface Page Read Times

The lengths of time between requests by expMi clients for interface pages, the apparent read times of interface pages, were calculated. These apparent read times were grouped into read durations of 1–2 seconds, 3–9 seconds, 10–29 seconds, 30–59 seconds, 59–300 seconds and 300+ seconds. The groupings for the four types of exploratory sessions are displayed in *Table 5-1*.

Chi square analysis on the apparent read times revealed some important findings. The p values are summarised in the following list:

- expMi defMi & expMi defEn; 0.85 p value, no significant difference
- expEn defMi & expEn defEn; 0.05 p value, significant difference
- expMi defMi & expEn defMi; 0.0006 p value, significant difference
- expMi defEn & expEn defEn; 0.60 p value, no significant difference

	expMi		expEn	
seconds	defMi	defEn	defMi	defEn
Total	133	98	221	215
1–2	29	23	21	35
3–9	20	13	55	37
10–29	28	25	70	64
30–59	17	15	29	26
59–300	22	13	34	30
300+	17	9	12	23

Table 5-1: defMi and defEn apparent read times of interface pages in expMi and expEn sessions.

The p values mentioned above indicate that clients who explored the Niupepa interface in reo Māori had similar apparent read times whether the default language was set to reo Māori or English. Also, the apparent read times of the expMi clients (for both default language settings) was similar to the apparent read times of the expEn clients when the default language of the web site was set to English. In contrast, the expEn clients exhibited significant statistical differences in apparent read times of interface pages with different settings of the default language.

Statistical summaries were calculated to determine what the differences were in apparent read times of Niupepa interface pages. These summaries are displayed in *Table 5-2*.

expMi and expEn Apparent Read Times for defMi and defEn				
	expMi		expEn	
	defMi	defEn	defMi	defEn
Total Apparent Reads	133	98	221	215
Average (seconds)	171.4	111.9	82.2	109.0
Standard Deviation	485.6	302.8	283.5	269.5
Mode (seconds)	0	0	9	0
Shortest session (seconds)	0	0	0	0
Lower Quartile (seconds)	4	4	8	6
Median (seconds)	21	18	16	18
Upper Quartile (seconds)	84	56	47	59
Longest session (seconds)	3396	3396	3453	1240

Table 5-2: Statistical summaries comparing apparent read times in expMi and expEn sessions with defMi and defEn settings.

High standard deviations indicate a wide variation in the apparent read times. Setting the default interface language to reo Māori corresponded with the greatest extremes in apparent read times. ExpMi clients had the longest average read time (171.4 seconds) when the default interface language was set to reo Māori. ExpEn clients had the shortest read times (82.2 seconds) when the default interface language was set to reo Māori.

Statistics describing the apparent read times of Niupepa interface pages in exploratory sessions reveal some important findings. The expMi page read times were not affected by the default language setting of the Niupepa interface and were similar to the expEn read times when the default language was English. The expEn apparent read times were significantly shorter when the default language was set to reo Māori.

5.2.2 Effect on Which Interface Pages Were Requested

The interface pages requested by exploratory reo Māori (expMi) clients in sessions with a default reo Māori setting (defMi) and in sessions with a default English language setting (defEn) are listed in *Table 5-3*. Also listed are the interface pages requested by exploratory English language (expEn) clients for the defMi and defEn settings of the Niupepa interface. Note that an interface page refers to a page of the Niupepa web site it does not refer to a newspaper document of the Niupepa collection.

Interface Pages Requested in defMi and defEn Exploratory Sessions								
Niupepa Interface Page Requested	expMi				expEn			
	defMi		defEn		defMi		defEn	
Total pages	232	100.0%	95	100.0%	399	100.0%	260	100.0%
niupepa home	192	82.8%	142	85.0%	336	84.2%	310	80.9%
NZDL home	9	3.9%	0	0.0%	30	7.5%	28	7.3%
About Eng. Abstr.	26	11.2%	11	6.6%	13	3.3%	14	3.7%
help	1	0.4%	3	1.8%	6	1.5%	15	3.9%
preferences	4	1.7%	10	6.0%	13	3.3%	15	3.9%
about greenstone	0	0.0%	0	0.0%	0	0.0%	1	0.3%
unknown	0	0.0%	1	0.6%	1	0.3%	0	0.0%

Table 5-3: Numbers and percentages of Niupepa interface pages requested by expMi and expEn clients with defMi and defEn language settings.

With chi square analysisⁱ a comparison of pages requested by expMi clients in defMi and defEn exploratory sessions generates a p value of $3.6 * 10^{-3}$. This indicates that there was a significant difference in the interface pages requested with the different default language settings.

When the default language was set to reo Māori, expMi clients requested more About English Abstracts pages. This could be caused by the ambiguous

ⁱ The requests numbers for ‘about greenstone pages’ and ‘unknown pages’ were not used when undertaking chi square analysis as the counts for these were too low.

translation of this hyperlink term (see §3.3.3 for further explanation). When the default language was set to English there were more requests for the preferences page; a likely scenario given that expMi clients were likely to switch the interface language from English to their preferred language of te reo Māori.

Using chi square analysis to compare the interface pages requested by clients who used English in defMi and defEn exploratory sessions generates a p value of 0.31. This indicates that there was no significant difference in the interface pages requested by expEn users when the default language was set to reo Māori compared with the interface pages requested when the default language was set to English.

5.2.3 Summary—Effect on Exploratory Behaviours

This section examined how the default language setting of the Niupepa interface affected te reo Māori clients who explored the Niupepa digital library but then did not go on to use it. The default language setting did not appear to have any effect on the duration (in time) and number of requests in exploratory sessions, either expMi or expEn. It also did not have any effect on the number of expMi clients who undertook exploratory sessions. However, setting the default language of the interface to reo Māori reduced the number of expEn clients by 28.6%.

The apparent read times of interface pages by expMi clients was similar for both defMi and defEn settings, and these read times were similar to read times taken by expEn clients with a defEn setting. However, apparent read times by expEn clients when the default language was set to reo Māori were significantly shorter.

There were significant differences in the pages requested by expMi clients using defMi and defEn settings. With a defMi setting, the About English Abstracts Page was requested more often; with a defEn setting, the Preferences page was requested more often. There were no significant differences in the Niupepa interface pages requested by expEn clients using the defMi and defEn settings.

5.3 Effect on Interface Behaviours

The previous section examined how the default language of the Niupepa interface affected behaviours by te reo Māori clients who explored the Niupepa digital library but then did not go on to use it. This section examines how the default language of the Niupepa interface affects behaviours by te reo Māori clients who actually did use the Niupepa digital library to request a Niupepa document and/or undertake a query activity.

There were 172,905 usage requests recorded in 6,714 sessions for the Niupepa collection in 2005. Of these, 87,678 (50.7%) requests were recorded in the 198 days that the default language of the web site was set to reo Māori (defMi) and 69,265 (40.1%) requests were recorded in the 137 days that the default language was set to English (defEn). These figures average out to 442.8 defMi requests per day and 505.6 defEn requests per day. The Niupepa digital library received 14% more usage requests when the default language of the web site was set to English. There were almost three times as many clients who preferred to use the interface in English.

Further, and perhaps not surprisingly, the number of usage session requests in a particular language increased when the default language of the web site was set to that language. With a defMi web site setting there were 19,628 (22.4%) requests in te reo Māori and 68,050 (77.6%) requests in English. With a defEn web site setting there were 11,217 (16.2%) requests in te reo Māori and 58,048 (83.8%) requests in English. This difference is illustrated in *Figure 5-3*.

The usage session requests were then further categorised into language session types. The counts for the various defMi and defEn usage session types were:

- usgMi defMi sessions; 92.2 requests/day (18,265 requests in 956 sessions)
- usgMi defEn sessions; 76.5 requests/day (10,475 requests in 623 sessions)
- usgEn defMi sessions; 337.8 requests/day (66,875 requests in 2,222 sessions)
- usgEn defEn sessions; 420.7 requests/day (57,630 requests in 2,322 sessions)

- usgBi defMi sessions; 12.8 requests/day (2,538 requests in 60 sessions)
- usgBi defEn sessions; 8.5 requests/day (3,698 requests in 26 sessions)

These figures show that when the default language of the interface was set to reo Māori, requests by clients who were using the interface in te reo Māori (usgMi) increased by 20.5%. When the default interface was set to English, requests by clients who were using the interface in English (usgEn) increased by 25.4%. With a defMi setting the ratio of usgMi to usgEn clients was 1:2.3; with a defEn setting the ratio of usgMi to usgEn clients was 1:3.8. Clearly, setting the default language of the interface to a particular language encouraged use of the interface in that language.

Bilingual usage sessions (usgBi) were 50.6% more likely to occur if the default language of the interface was set to reo Māori.

Five interface behaviours were further investigated to determine if the default language setting of the Niupepa digital library affected how the interface was used in te reo Māori by Niupepa clients. The interface behaviours investigated were:

- the length of defMi and defEn usgMi sessions
- the apparent read times of interface pages in defMi/defEn usgMi sessions
- the number of return defMi and defEn usgMi sessions
- the interface pages requested in defMi and defEn usgMi sessions
- the interface language switching in defMi and defEn usgMi sessions.

These five interface behaviours are described in the next five sub-sections.

5.3.1 Effect on Usage Session Length

Two measurements are used to define session lengths; duration in time (measured in minutes) and number of requests. Statistics derived from TLA were used to compare lengths of usage sessions when the default language was set to reo Māori and set to English. These comparisons are detailed in the next two sub-sections.

Usage Session Duration in Time

The durations of the usage sessions were grouped into time intervals of less than 1 minute, 1–2 minutes, 3–9 minutes, 10–29 minutes, 30–59 minutes and 60+ minutes. The number of sessions that fitted into these time categories was counted for the different language types of defMi and defEn usage. The counts and the p values of chi square analysis on these counts are displayed in *Table 5-4*.

Duration Groupings of Usage Sessions (in minutes)						
time (minutes)	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
<1	207	224	289	710	0	0
1–2	200	102	523	434	2	2
3–9	185	105	460	352	7	3
10–29	189	103	425	412	22	6
30–59	105	49	275	225	19	10
60+	70	40	250	189	10	5
Total	956	623	2222	2322	60	26
p values	1.5 * 10⁻⁰⁷		9.2 * 10⁻⁴⁴		0.70	

Table 5-4: defMi and defEn comparisons for length of session groupings (in minutes) of usage sessions.

A p value of $1.5 * 10^{-07}$ indicates a significant statistical difference in usage session lengths of usgMi defMi and usgMi defEn sessions. The default language setting affected the duration of usage sessions undertaken by clients who used the Niupepa interface in te reo Māori.

Similarly, with a p value of $9.2 * 10^{-44}$ the default language setting also affected the duration of usage sessions undertaken by clients who used the Niupepa interface in English. By contrast, the length of usage sessions by clients who used the Niupepa interface bilingually was not affected by the default language setting (p value 0.7).

Statistical summaries of session lengths were calculated to determine where the defMi and defEn session duration differences were. These statistical summaries are displayed in *Table 5-5*.

Summary of defMi and defEn Usage Sessions Lengths (Minutes)						
Minutes per session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Average	16.8	15.5	22.9	17.9	44.1	42.7
Standard Deviation	27.3	34.7	36.6	34.8	50.2	42.3
Most Common	0	0	0	0	21	21
Shortest Session	0	0	0	0	3	3
Lower Quartile	0	0	2	0	16	21
Median	5	3	7	4	25	35.5
Upper Quartile	20	14	29	19	48	52
Longest Session	333	321	319	356	256	219

Table 5-5: *defMi and defEn statistical summaries for duration (in minutes) of usgMi, usgEn and usgBi usage sessions.*

The statistics displayed in *Table 5-5* show that the mean, the median and the upper quartile of usgMi sessions were higher with a defMi setting than with a defEn setting. Consequently, usgMi sessions had a longer duration when the default language of the interface was set to reo Māori. Similarly, usgEn sessions also had a longer duration with a defMi setting.

Usage Sessions Length in Number of Requests

Usage session length analysis was also undertaken on the number of requests submitted per usage session. Requests were collated into request counts per session of 2 requests, 3–4 requests, 5–8 requests, 9–19 requests, 20–39 requests and 40+ requests. The counts and the chi square p values are shown in *Table 5-6*.

Length in Requests of defMi and defEn Usage Sessions						
Requests per session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
2	106	123	107	294	n/a	n/a
3–4	108	94	151	343	n/a	n/a
5–8	247	137	540	478	n/a	n/a
9–19	222	132	545	483	11	8
20–39	147	80	402	337	26	9
40+	126	57	477	387	23	9
Total	956	623	2222	2322	60	26
p values	$2.7 * 10^{-06}$		$1.7 * 10^{-37}$		0.43	

Table 5-6: defMi and defEn request groupings with p values.

A p value of $2.7 * 10^{-06}$ indicates a significant statistical difference in the number of requests in usgMi defMi and usgMi defEn sessions. Consequently the default language setting did affect the number of requests per sessions submitted by clients who used the Niupepa interface in te reo Māori.

A p value of $1.7 * 10^{-37}$ indicates a significant statistical difference in the number of requests in usgEn defMi and usgEn defEn sessions. Statistical summaries were calculated to display the differences in numbers of requests and these summaries are displayed in *table 5-7*.

The Niupepa default language setting affected the number of requests that were submitted by clients who used the interface in te reo Māori. When the default language of the interface was set to reo Māori, more requests were submitted by usgMi clients than when the default language of the interface is set to English. Although this difference did not appear to be much, an average of 19.1 requests compared to an average of 16.8 requests, chi square analysis did show the difference to be statistically significant.

Requests per Session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Average	19.1	16.8	30.1	24.8	42.3	44.6
Standard Deviation	24.5	27.5	42.4	41.8	30.7	45.9
Mode	2	2	6	2	21	30
Fewest requests	2	2	2	2	10	14
Lower Quartile	5	3	6	4	22	19
Median	10	7	15	10	30.5	30
Upper Quartile	23	18	35	28	53	53
Most requests	218	245	598	463	168	239

Table 5-7: *defMi and defEn comparisons of statistical summaries for length (in requests) of usgMi, usgEn and usgBi usage sessions.*

Clients who used the interface in English also submitted significantly more requests when the default interface language was set to reo Māori. Bilingual clients did not show a significant difference in the number of requests submitted per session for the two default language settings of the Niupepa web site.

5.3.2 Effect on Time to Read Interface Pages

As mentioned in §2.5.1, an apparent page read time is the length of time from when one page is requested until another page is requested (where a page refers to a page of the interface as opposed to a document from the collection). The apparent read times can be compared for the different default language settings.

Statistical summaries for defMi and defEn apparent read times are listed in *Table 5-8*.

Apparent Read Times of Interface Pages in defMi and defEn Sessions						
Exploratory Sessions:	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Apparent Reads	797	415	2873	2363	304	196
Average (seconds)	44.7	49.1	48.5	45.6	59.9	98.2
Standard Deviation	138.3	151.9	185.8	168.7	178.7	292.8
Mode (seconds)	4	4	5	4	4	3
Shortest (seconds)	0	0	0	0	0	0
Lower Quartile (seconds)	5	5	6	6	6	6
Median (seconds)	10	10	13	12	14	18.5
Upper Quartile (seconds)	29	34	33	31	35	51
Longest (seconds)	1826	2131	2874	3237	2041	2367

Table 5-8: defMi and defEn statistical comparisons of apparent read times of Niupepa interface pages in usgMi, usgEn and usgBi usage sessions.

An initial look at the average read times above would tend to suggest that clients who used the interface in te reo Māori (usgMi) spent less time reading the interface pages when the default language of the web site was set to reo Māori (44.7 seconds compared with 49.1 seconds). Similarly, the figures tend to suggest that usgEn clients spent less time reading the interface pages with a defEn setting (average 45.6 seconds) than with a defMi setting (48.5 seconds).

However a chi square test on the apparent read times by usgMi clients in defMi and defEn sessions generates a p value of 0.75. This value indicates that there was no statistically significant difference in the apparent read times of usgMi clients using the different default language settings of the Niupepa interface. With p values of 0.03 and 0.14 there was also no statistically significant difference in apparent read times of usgEn clients or usgBi clients using the different default interface settings.

The default language setting of the Niupepa interface did not affect how long usgMi clients took to read interface pages of the Niupepa digital library.

5.3.3 Effect on Return Requests

Return requests are requests submitted by clients who have used the Niupepa digital library previously. In 2005 an average of 11.8 return requests were recorded, per day, by usgMi clients using the defMi interface setting. There was also an average of 12.4 return requests recorded, per day, by usgMi clients using the defEn setting. This difference is not significant. The numbers of return clients who used the interface in English did not show a significant difference which could be attributed to the default language setting; there were, on average, 73.4 return requests per day with a defMi setting and 72.5 return requests per day with a defEn setting.

The number of return requests by clients who used the Niupepa digital library in te reo Māori (and English) was not influenced by the default language setting of the Niupepa interface.

5.3.4 Effect on Which Interface Pages Were Requested

The percentages of interface pages requested in usgMi sessions with defMi and defEn settings are compared in *Table 5-9*. Also listed are the percentages of interface pages requested in usgEn and usgBi sessions.

Table 5-9 suggests that altering the default language of the web interface did not affect which interface pages were requested by usgMi clients. The percentages of usgMi pages requested were similar for both the defMi and defEn setting. A chi square test (with the non-requested pages removed) confirms this, with a p value of 0.14. Thus, the default language of the Niupepa digital library did not influence which pages were requested by usgMi clients.

Interface Pages Requested in defMi and defEn Usage Sessions						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total pages	827	432	3,005	2,476	307	202
Niupepa home page	80.5%	81.3%	83.4%	79.6%	84.0%	63.4%
NZDL home page	1.7%	0.5%	1.4%	3.0%	2.9%	0.0%
About English Abstract	6.7%	4.6%	5.6%	5.8%	4.9%	3.5%
help page	3.3%	4.4%	2.1%	2.6%	1.6%	7.4%
preferences page	7.9%	9.3%	7.6%	9.0%	6.5%	25.7%
about greenstone page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
unknown page	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
pages per day	4.2	3.2	15.2	18.1	1.6	1.5

Table 5-9: defMi and defEn comparison of percentages of Niupepa web site interface pages requested in usgMi, usgEn and usgBi sessions.

However, the default language of the Niupepa digital library did appear to affect the interface pages requested by usgEn clients and usgBi clients. Both client categories showed different percentages of pages requested for the different default language settings. With p values of 6.7×10^{-5} and 2.4×10^{-11} these differences were statistically significant. The default language of the Niupepa digital library did influence which pages were requested by usgEn and usgBi clients.

5.3.5 Effect on Interface Language Switching

The number of times a client switched the language of an interface per week was recorded and compared for defMi and defEn web interface settings. These numbers are displayed in *Table 5-10*.

defMi and defEn Web Site Interface Language Switching Summaries						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Switches	102	74	753	390	209	124
Māori to English/week	1.6	1.4	23.3	16.1	4.4	3.3
English to Māori/week	2.0	2.4	3.4	3.8	3.0	3.1
Total Switches/week	3.6	3.8	26.7	19.9	7.4	6.4
Switches/Requests	0.56%	0.71%	1.13%	0.68%	8.23%	3.63%

Table 5-10: defMi and defEn total interface switches, switches per week and total switches/total requests for all usage sessions types.

It is apparent from *Table 5-10* that switching the interface language of the Niupepa digital library occurred infrequently. The 176 switches (102 +74) that occurred over all the usgMi sessions account for 0.6% of all usgMi requests. The 1,140 switches (753+390) that occurred in all usgEn sessions account for 0.9% of all requests in usgEn sessions. The 333 (209+124) switches that occurred in all usgBi sessions account for 5.6% of all usgBi requests. However, by definition bilingual sessions must include at least three language switches.

As might be expected, there were higher proportion of reo Māori to English switches when the default language was set to te reo Māori and a higher proportion of English to reo Māori switches when the default language of the interface was set to English. These higher proportions of switches were behaviour displayed by all client types usgMi, usgEn and usgBi clients.

The switching data was also collated by the number of interface language switches that occurred per usage session. This data is displayed in *Table 5-11*. Most of the switches in usgMi and usgEn sessions consisted of just one or two switches per session. Almost half of the usgMi session switching consisted of just one switch per session, with the reminder consisting of two switches per session. By definition the bilingual sessions consisted of only 3+ switches per session.

defMi and defEn Web Site Interface Language Switches per Session						
Switches per session	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
1 switch	48.0%	48.6%	75.6%	66.4%	n/a	n/a
2 switches	49.0%	43.2%	21.5%	32.8%	n/a	n/a
3+ switches	2.9%	8.1%	2.9%	0.8%	100.0%	100.0%
	100%	100%	100%	100%	100%	100%
Switch Sessions/ Total Sessions	7.85%	8.67%	29.57%	13.95%	100%	100%

Table 5-11: defMi and defEn switches per session for all usage types.

Did the default language of the Niupepa digital library influence the language switching undertaken by usgMi clients? No, not significantly. While there were minor increases of language switching from reo Māori to English with a defMi setting, and English to reo Māori with a defEn setting, overall these were not statistically significant. *Table 5-10* shows a difference of just 0.2 total switches per week (3.6 compared with 3.8) for the different defMi and defEn settings. *Table 5-11* shows a difference of just 0.82% (7.85% compared with 8.67%) of switch sessions per total sessions for the different defMi and defEn settings. UsgMi clients, unlike usgEn clients, switched the language of the interface no matter what the default language was set to.

5.3.6 Summary—Effect on Interface Behaviours

This section examined how the default language of the Niupepa digital library affected behaviours by te reo Māori clients who used the Niupepa interface. Statistics derived from TLA were used to show how the default language setting affected the number of requests to the Niupepa digital library and five particular aspects of how usgMi clients used the Niupepa interface.

The statistics revealed that there were 21% more requests in te reo Māori when the default language of the interface was set to reo Māori (defMi) and 25% more

requests in English when the default language of the interface was set to English (defEn). With a defMi setting the ratio of usgMi clients to usgEn clients was 1:2.3, with a defEn setting the ratio of usgMi clients to usgEn clients was 1:3.8. Setting the default language of the Niupepa interface to reo Māori resulted in more requests being undertaken in te reo Māori and in a higher ratio of usgMi to usgEn clients.

As mentioned above, statistics from TLA were used to examine five aspects of usgMi client behaviour when using the Niupepa interface. These statistics showed that sessions by usgMi clients had a longer duration and consisted of more requests if the default language of the interface was set to reo Māori. None of the four other aspects examined was affected by the default language setting of the interface. Clients who used the interface in te reo Māori had similar apparent page read times, similar numbers of return users, requested similar interface pages and showed similar language switching behaviours whether the default language of the interface was set to reo Māori or to English.

This summary concludes the examination into how the default language setting of the Niupepa interface affected usage of the interface by usgMi clients. The next aspect to be examined is whether the default language setting affected how te reo Māori clients accessed documents in the Niupepa digital library.

5.4 Effect on Document Access Behaviours

Did the default language affect how usgMi clients accessed documents in the Niupepa digital library? This question is addressed by examining four aspects of document access:

- how the documents were accessed
- which formats were accessed
- which document types were accessed
- the apparent document read times of the documents that were accessed.

5.4.1 Effect on How Documents Were Accessed

Three methods were available to clients accessing Niupepa documents. Documents were able to be selected from a list of results following a query activity, a list of publication series and a list of articles published in a certain date. Statistics from TLA indicated the methods used by usgMi clients when the default language of the interface was set to reo Māori and to English. These statistics are displayed in *Table 5-12*.

How Documents Were Requested With defMi and defEn Settings						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Documents	13,030	7,427	49,533	42,709	1,539	609
From query	77.6%	76.3%	83.9%	78.0%	60.8%	90.1%
From series	14.4%	10.9%	4.3%	9.9%	13.5%	3.6%
From date	6.1%	10.7%	9.6%	9.9%	21.1%	3.4%
Uncertain	1.8%	2.1%	2.2%	2.3%	4.7%	2.8%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Documents per day	65.8	54.2	250.2	311.7	7.8	4.4
p value	3.2×10^{-37}		2.8×10^{-241}		3.5×10^{-39}	

Table 5-12: *defMi and defEn comparison by percentage of how documents were requested by usgMi, usgEn and usgBi clients.*

Using the search facility to submit a query was the most common method of accessing documents. This was true for all client types and for both default interface language settings. This preference was true for both defMi (77.6%) and defEn (76.3%) settings.

However, the p value of 3.2×10^{-37} indicates that usgMi clients were accessing documents in a significantly different manner in each of the default language settings. The significant difference showed in the preference for the series and date lists; when the default language was set to reo Māori usgMi clients displayed a stronger preference for using the series lists (14.4% compared with 6.1%). When

the default language was set to English this preference was less pronounced and usgMi clients displayed an almost equal preference for using the series and date lists (10.9% compared with 10.7%).

The default language of the web site also affected browsing strategies undertaken by usgEn clients. A defMi setting seemed to encourage usgEn clients to use the query facility (83.9% compared with 78.0%) and seemed to discourage the selection of documents from the series list (4.3% compared with 9.9%).

The default language of the web site seems to have had the most effect on the information retrieval strategies of usgBi clients. When the default language was set to reo Māori they submitted a third fewer queries (60.8% compared with 90.1%), the series list was used almost 10% more (13.5% from 3.6%) and the date list was used 17.7% more (21.2% from 3.4%). Setting the default language to reo Māori seemed to encourage the usgBi clients to be more adventurous in their information retrieval strategies.

Thus, the default language setting of the Niupepa web site significantly affected how Niupepa documents were accessed by all client types.

5.4.2 Effect on Document Formats Requested

As mentioned in §3.1, three document formats were available to the client. *Table 5-13* lists the percentages of requests for each of the formats by usgMi clients using both the defMi and defEn settings. Also listed are the percentages requested by usgEn and usgBi clients.

Formats of Requested Documents With defMi and defEn Settings						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total Documents	13,030	7,427	49,533	42,709	1,539	609
Extracted Text	65.3%	53.4%	49.8%	62.8%	58.3%	43.0%
Full Size Image	29.8%	40.6%	44.7%	29.1%	36.5%	35.5%
Preview Image	4.1%	5.8%	5.2%	7.3%	4.9%	21.0%
Undefined	0.9%	0.1%	0.3%	0.8%	0.3%	0.5%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
p value	1.9 x 10 ⁻⁷⁴		0		8.0 * 10 ⁻³⁰	

Table 5-13: *defMi and defEn comparisons of document types requested by usgMi, usgEn and usgBi clients.*

The p values listed in *Table 5-13* indicate that the default language setting of the interface had a statistically significant impact on which document formats were requested by clients of the Niupepa digital library. The extracted text format was the favoured format by all client types, with both defMi and defEn settings; however, it seemed to be less favoured when the default language did not match the usage language of the clients.

For usgMi clients the extracted text documents were requested more (65.3% compared with 53.4%) when the default language of the interface was set to reo Māori. When the default language of the interface was set to English the full size images were requested at a higher rate (40.6% compared with 29.8%).

Similarly, the usgEn clients requested a higher percentage (62.85 compared with 49.8%) of extracted text documents when the default language of the interface was set to English. When the default language of the interface was set to reo Māori the full size images were preferred at a higher rate (44.7% compared with 29.1%) by usgEn clients.

5.4.3 Effect on Which Document Types Were Requested

Five different types of documents were available to Niupepa clients. These documents consisted of; newspapers in te reo Māori, newspapers in English, parallel (reo Māori/English) newspapers, abstracts in English and commentary information written in both English and reo Māori (see §3.5.3). The types of documents requested when the default language of the interface was set to reo Māori were compared with those requested when the default language was set to English; significant differences were noted.

Rather than simply counting the document types requested, the document types requested by unique users were counted. Unique user counts give a better picture of what a wider range of people accessed and limits the influence in the statistics of the heavy users. It was also decided to collate only the top 100 unique pages requested, as limiting the analysis to this number gives a better indication of users' initial preferences and, again, reduces the influence in the statistics of heavy users of the web site.

Documents requested by usgBi clients were excluded as the number of requests submitted by usgBi clients were considered too low to generate meaningful results. The data collated for this analysis consisted of:

- 774 unique user requests for 100 pages in usgMi sessions with defMi setting
- 686 unique user requests for 141 pages in usgMi sessions with defEn setting
- 2048 unique user requests for 98 pages in usgEn sessions with defMi setting
- 2185 unique user requests for 110 pages in usgEn session with defEn setting

Requests for each document type are displayed as percentages in *Figure 5-7*.

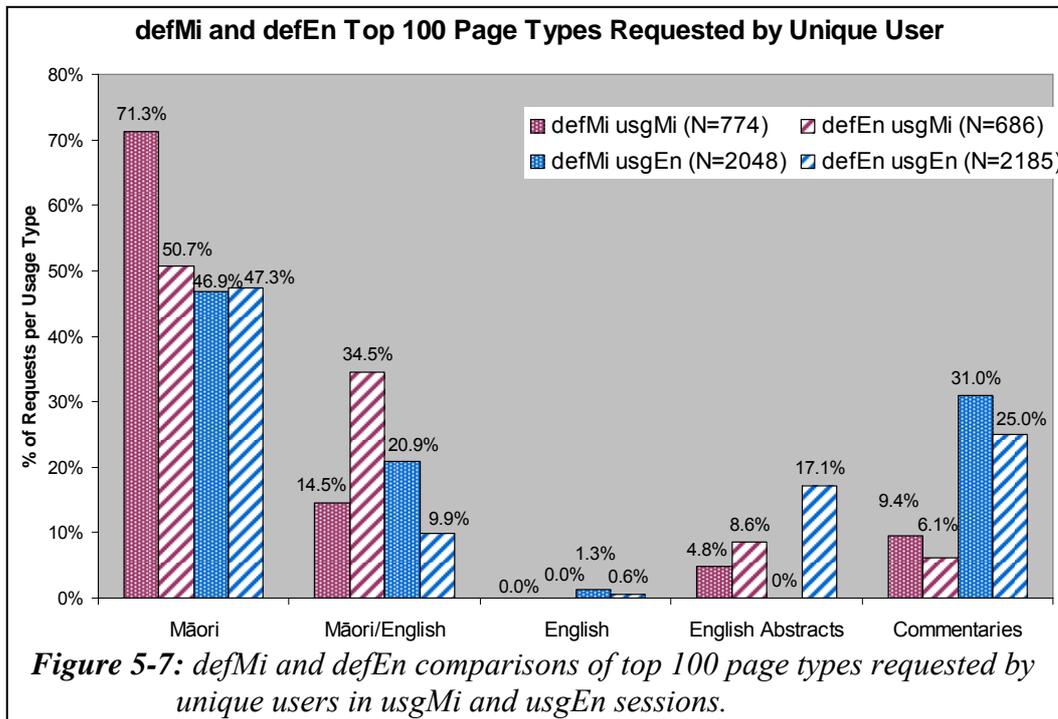


Figure 5-7 shows that the default language of the Niupepa interface did affect the type of documents accessed by usgMi clients. When the default language was set to reo Māori, usgMi clients requested reo Māori newspapers in 71.3% of their requests and parallel (reo Māori/English) newspapers in only 14.5% of requests. However, when the default language was set to English, usgMi clients requested reo Māori newspapers in only 50.7% of their requests and parallel newspapers in 34.5% of requests. UsgMi clients were twice as likely to request English Abstracts (8.6% compared with 4.8%) when the default language was English.

Figure 5-7 also highlights an unusual behaviour by usgEn clients. The English Abstracts were requested in 17.1% of requests when the default language of the web site was set to English, but, incredibly, weren't requested at all when the default language of the web site was set to reo Māori. It is difficult to imagine why a usgEn client would not want to select an English Abstract simply because the default language of the web site was set to reo Māori. Perhaps with the interface initially set to reo Māori, the usgEn clients were not aware that English Abstracts were available.

5.4.4 Effect on Reading Time Of Documents

The apparent document read times (as explained in §2.5.1) by usgMi clients were compared for the defMi and defEn settings. The average apparent read times of full size documents, of extracted text documents and of preview documents were calculated and are displayed in *Table 5-14*.

Average Apparent read times (seconds)	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
Total	70.9	73.9	56.2	53.0	77.8	57.9
Full Size Image	117.7	106.1	90.7	96.6	135.3	86.1
Extracted Text	52.3	54.9	29.1	36.9	47.0	53.1
Preview Image	43.4	23.0	26.7	25.5	27.9	22.2

Table 5-14: defMi and defEn comparisons of average apparent read times (in seconds) of documents by usgMi, usgEn and usgBi clients.

These figures indicate that when the default language was set to reo Māori (defMi) the usgMi clients took longer, on average, to read the full size images (117.7 seconds compared with 106.1 seconds) and the preview images (43.4 seconds compared with 23.0 seconds). However, there was no significance difference in the apparent document read times of extracted text. The default language setting did affect the apparent read times of preview and full size images but not extracted texts.

5.4.5 Summary—Effect on Document Access

This section examined how the default language setting of the Niupepa interface affected how te reo Māori clients accessed documents. Statistics produced from TLA were used to reveal significant differences in four document access behaviours.

Using a query was the method preferred by Niupepa clients to access documents. Statistics showed that 75% of the usgMi requests for documents resulted from a query, a rate consistent for both default interface language settings. However, usgMi clients showed a higher preference for using the series list when the default language was set to reo Māori than they did when the default was set to English.

Extracted text was the preferred document format in 56% of Niupepa requests. There were 12% more requests by usgMi clients for the extracted text format with a defMi setting. Alternatively, there were 11% more requests by usgMi clients for the full size image format with a defEn setting.

The default language setting of the Niupepa interface also affected the type of documents requested by usgMi clients. UsgMi clients requested significantly more reo Māori newspapers and significantly fewer parallel (reo Māori/English) newspapers when the default interface language was set to reo Māori. UsgMi clients also requested more English Abstracts when the default interface language was set to English.

The apparent read times of extracted text documents were similar for both the defMi and defEn settings. However, usgMi clients took longer to read full size images and preview images when the default language was set to reo Māori.

This section analysed how the Niupepa default language setting affected document access behaviours of te reo Māori clients. The third aspect of behavioural analysis, query behaviours, is examined in the next section.

5.5 Effect on Query Behaviours

This section analyses how the default language setting affected the query behaviours of clients who used the Niupepa digital library in te reo Māori. A query is defined as using the search facility of the Niupepa digital library to find a term or string of terms in documents stored in the Niupepa digital library.

The TLA process generated query behaviour statistics on the data collected when the Niupepa default language was set to reo Māori (defMi) and to English (defEn). Five query behaviours of usgMi clients are examined in this section:

- the number of queries submitted
- the number of query terms used
- the language of query terms
- the use of query parameters
- the use of query operators.

5.5.1 Effect on Number of Queries

The numbers of queries submitted per session by usgMi clients when the default language was set to reo Māori were compared with those submitted when the default was set to English. Percentages were calculated and are listed in *Table 5-15* for comparison. Also displayed are the percentages of queries per session by usgEn and usgBi clients for the defMi and defEn settings.

Table 5-15 shows that the default language setting of the Niupepa interface did affect the number of queries per usgMi session. UsgMi sessions were more likely to involve a query when the Niupepa interface was set to defMi.

	usgMi		usgEn		usgBi	
Queries/Session	defMi	defEn	defMi	defEn	defMi	defEn
no queries	39.4%	50.6%	27.0%	41.1%	20.0%	23.1%
1–2 queries	32.5%	22.5%	35.0%	28.6%	13.3%	11.5%
3+ queries	28.0%	27.0%	38.0%	30.3%	66.7%	65.4%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Total Sessions	956	623	2222	2322	60	26

Table 5-15: defMi and defEn comparisons of number of sessions that contained queries by usgMi, usgEn and usgBi clients.

It appears that approximately 11% of the usgMi sessions were influenced by the defMi and defEn setting. With a defMi setting 60.5% (32.5% + 28.0%) of the usgMi sessions involved the submission of at least one query, with a defEn setting, only 49.5% (22.5% + 27.0%) of the usgMi sessions involved the submission of at least one query.

The default language setting also affected the number of queries that were submitted per usgEn and per usgBi session. Sessions undertaken by usgEn and usgBi clients showed a significantly higher rate of queries per session when the default language of the interface was set to reo Māori. Both usgEn and usgBi clients had a higher percentage of queries submitted per sessions than usgMi clients.

5.5.2 Effect on Number of Query Terms

The numbers of query terms in queries submitted by usgMi clients with a defMi setting were compared with those submitted by usgMi clients with a defEn setting. The percentages of these terms were calculated and are displayed in *Table 5-16*. Also displayed are the numbers of usgEn and usgBi query terms for the different default language settings of the Niupepa interface.

	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
1 term/query	45.6%	48.1%	47.8%	51.2%	43.9%	52.5%
2 terms/query	28.8%	31.9%	33.3%	30.8%	39.4%	24.6%
3+ terms/Query	25.7%	20.0%	18.9%	17.9%	16.6%	23.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Queries	2666	1703	8650	6871	289	183
p value	$8.1 * 10^{-5}$		$1.0 * 10^{-4}$		$3.3 * 10^{-3}$	

Table 5-16: defMi and defEn percentage comparisons for the number of terms in a query in usgMi, usgEn and usgBi sessions.

The default language setting affected the number of query terms used per query by usgMi clients. With a p value of $8.1 * 10^{-5}$ the usgMi sessions showed a statistically significant difference in the number of query terms used for the defMi and defEn settings. The percentage of usgMi queries with just one term was lower with the defMi setting (45.6% compared with 48.1%). The percentage of usgMi queries with 3+ query terms was higher for the defMi setting (25.7% compared with 20.0%). These percentages indicate that with a defMi setting usgMi clients submitted more query terms per query.

The default language setting also affected the number of query terms used per query by usgEn and usgBi clients; they also entered more query terms per query when the default language of the interface was te reo Māori.

5.5.3 Effect on Language of Query Terms

The language of the query terms most requested by usgMi clients were examined to see if the default language setting had an influence over which language was preferred. These consisted of 116 terms submitted 479 times when the default language was set to te reo Māori and 85 terms submitted 352 times when the default language was set to English. The percentages of query terms in te reo Māori and in English are displayed in *Table 5-17*.

The default language setting of the Niupepa interface did affect the language of query terms used by usgMi clients. With a defMi setting, 94.4% of their query terms were entered in te reo Māori. With a defEn setting only 86.1% of their query terms were entered in te reo Māori. Setting the default language to te reo Māori corresponded with a higher number of query terms submitted in te reo Māori.

Language of Query Terms With defMi and defEn Settings						
	usgMi		usgEn		usgBi	
	defMi	defEn	defMi	defEn	defMi	defEn
reo Māori terms	94.4%	86.1%	83.3%	83.0%	74.8%	74.7%
English terms	5.6%	13.9%	16.7%	17.0%	25.2%	25.3%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
Terms Examined	479	352	965	962	163	83
Unique Terms	116	85	104	118	62	30

Table 5-17: defMi and defEn percentage comparisons for language of query terms submitted by usgMi, usgEn and usgBi clients.

Interestingly, the default language of the Niupepa interface did *not* affect the language of the query terms used by usgEn and usgBi clients. UsgEn clients showed similar percentages of query terms in te reo Māori for both default language settings (83.8% compared with 83.0%). UsgBi clients also showed similar percentages of query terms in te reo Māori for both default language settings (74.8% compared with 74.7%).

5.5.4 Effect on Altering of Query Parameters

Earlier statistics (see §3.6.5) indicated that the only query parameter to be altered significantly by Niupepa clients was the Some/All parameter. With a defMi setting 7.0% of usgMi clients altered this parameter, while with a defEn setting 9.7% altered this parameter. This difference was not significant enough to conclude that the default language affected how much this query parameter was used by usgMi clients.

The default language setting did seem to affect how often the Some/All query parameter was used by usgEn clients. With a defMi setting, 12.1% of usgEn clients altered this parameter, while with a defEn setting only 8.8% altered this

parameter. It appears that if usgEn clients were prepared to switch the language of the interface from reo Māori to English they were also more likely to alter the Some/All query parameter.

5.5.5 Effect on Use of Query Operators

Earlier statistics (see §3.6.6) indicated that the only query operators used significantly by Niupepa clients were quotation marks; those were used to define precise phrase querying. With a defMi setting, 16.6% of queries submitted by usgMi clients included the use of quotation marks, while with a defEn setting 17.1% of queries included quotation marks. Thus, the default language setting of the Niupepa web site did not affect the number of query operators used by usgMi clients.

The default language setting did appear to affect the number of usgEn clients who used the quotation marks to define precise phrase querying. With a defMi setting, 4.9% of queries submitted by usgEn clients included the use of quotation marks. This changed to 8.9% of queries with a defEn setting. UsgMi clients were only half as likely to use quotation marks if the default language of the interface was set to reo Māori.

5.5.6 Summary—Effect on Query Behaviours

This section analysed how the default language setting affected the query behaviours of clients who used the Niupepa digital library in te reo Māori. Statistics produced from TLA were used to examine five query behaviours of usgMi clients and identify differences arising from the different default language settings of the Niupepa interface.

The statistics revealed that the default language setting of the Niupepa interface significantly affected three query behaviours of usgMi clients. When the default

language of the interface was set to reo Māori usgMi clients were more likely to submit a query, used more query terms per query and were more likely to enter the terms in te reo Māori.

The default language setting of the Niupepa interface did not appear to affect how often the Some/All query parameter was used by usgMi clients or the use of quotation marks to define precise phrase matching.

Analysis in the last three sections of this chapter has revealed that setting the default language to reo Māori encouraged the use of the Niupepa digital library in te reo Māori. With a defMi setting there was more use of the interface in te reo Māori, more te reo Māori terms when querying and documents written in te reo Māori were requested more. Before this chapter is concluded a further analysis merits discussion—how the default language setting affected clients who used the Niupepa digital library in English. This question is addressed in the following section.

5.6 Effect on Use in English

The previous three sections have examined how the default language setting affected the usage of the Niupepa digital library by clients who predominantly used the interface in te reo Māori (usgMi). Those sections showed that use of te reo Māori in almost all aspects of information retrieval increased when the default language of the interface was set to reo Māori. However, given that there were almost three times as many usgEn as usgMi clients, a further question needs to be considered. How did setting the default language to reo Māori affect usage of the Niupepa digital library by clients who predominately used the interface in English?

Setting the default language to reo Māori reduced the total number of filtered requests. Statistics from TLA revealed that 12.5% fewer Niupepa requests were received with a defMi setting. The number of requests by usgEn clients was

reduced even more as 6.2% of the usgEn requests had shifted to be recorded as usgMi requests.

While the number of usgMi clients remained constant for both default language settings, this was not the case for the usgEn clients. There were 1/3 (36.2%) fewer usgEn clients when the default language of the interface was set to reo Māori.

With a defMi setting, usgEn clients undertook longer usage sessions (both in duration and number of requests), requested the Niupepa home page at a higher rate and switched the interface language to English more often. These behaviours were perhaps a consequence of the default language setting not matching the clients' preferred language of use and, thus, these clients had to undertake the extra step of switching the interface language into a language that they could use.

When the default language was set to reo Māori, usgEn clients were more reliant on the query facility to access documents and had a higher preference for requesting full size images. One surprising aspect was that usgEn clients did not request English Abstracts when the default language of the interface was set to reo Māori. This behaviour is perhaps a usability awareness issue with the Niupepa web site as English Abstracts were created as a resource for usgEn clients no matter what the setting of the default language was.

UsgEn clients submitted more queries per session during a defMi setting. They also altered the Some/All query parameter more often and submitted more query terms per query with this setting. A higher number of query terms per query often means more use of the quotation marks to define phrase matching; however, this operator was not used by usgEn clients as much with a defMi setting.

In general, when the default language was set to reo Māori, usgEn clients used fewer features of the Niupepa interface and were more reliant on the query facility to access documents.

It is difficult to confirm that the behavioural differences noted by the usgEn clients with the different default language settings were solely a result of the default language setting of the Niupepa interface. The behaviours displayed by the

usgEn clients with a defMi setting could be the behaviours that these clients undertook for both default language settings. However, these behaviours only become apparent when the 1/3 of usgEn clients intolerant of the defMi setting were not included in the statistics.

What can be confirmed is that 1/3 of the usgEn clients were unwilling to use the Niupepa digital library if the default language was set to reo Māori, resulting in 12.5% fewer requests to the Niupepa digital library when the interface was set to te reo Māori.

5.7 Summary

This chapter has continued the investigation into how indigenous languages are used by clients of digital libraries. In particular, this chapter examined how the default language setting of the Niupepa digital library affected te reo Māori usage of the Niupepa digital library. The results have shown that setting the default language to reo Māori significantly increased the number of usgMi requests and, perhaps more importantly, increased the number of situations where te reo Māori was used to undertake information retrieval activities.

Setting the default language to reo Māori led to a 21% increase in reo Māori usage requests to the Niupepa digital library. This increase was mostly from clients who were already using the digital library in te reo Māori.

Setting the default language to reo Māori had very little effect on how the Niupepa digital library was explored (but then not used) by clients who used te reo Māori. The number of te reo Māori exploratory clients, the duration of te reo exploratory sessions, the number of requests in reo Māori exploratory sessions and the apparent interface page read times in reo Māori exploratory sessions were all similar for both default language settings. The only difference noted was the number of reo Māori exploratory clients who accessed the About English

Abstracts button; a difference most likely caused by an ambiguity issue arising from an inaccurate translation of the button label.

Setting the Niupepa default language to reo Māori did affect the length of reo Māori usage sessions; they were of a longer duration and consisted of more requests. However, other usgMi interface characteristics (such as apparent page read times, numbers of return users, interface pages requested and language switching behaviours) were unaffected by whether the default language was set to reo Māori or English.

The manner in which te reo Māori clients accessed documents was affected by the default language setting. These clients showed a stronger preference for using the series list, accessing extracted text formats, accessing Niupepa documents written in te reo Māori, and took longer to read full size and preview images when the default language was set to reo Māori.

Some query behaviours of te reo Māori clients were affected by the default language setting of the Niupepa web site. UsgMi clients were more likely to submit a query, used more query terms per query and were more likely to enter the terms in te reo Māori when the default language of the interface was set to reo Māori. However, use of the Some/All query parameter and the quotation marks query operator were not affected by the default language setting of the Niupepa web site.

The default language setting of the Niupepa interface also affected clients who used the interface in English. Setting the default language of the interface to reo Māori reduced the number of English exploratory (expEn) requests by 28.6% and the number of English usage (usgEn) requests by 19.7%. When the default language of the interface was set to reo Māori clients who used the digital library in English used fewer features of the Niupepa interface and were more reliant on the query facility to access documents.

5.7.1 Discussion

Data from the 2006 census (Statistics New Zealand, 2006) indicates that 98.5% of Māori are able to speak English. Developers of a reo Māori/English bilingual information system could, perhaps, assume that English would be a suitable default language as it would be useable for both English speaking and reo Māori bilingual clients. However, setting the default language of a bilingual system to English would not assist in the revitalisation of te reo Māori. In fact, it would do the opposite. Statistics derived from TLA on the Niupepa digital library in 2005 have shown te reo Māori was used significantly less by te reo Māori clients when the default language was not set to reo Māori.

Alternatively, setting the default language of a bilingual interface to an indigenous language encourages use of the interface in the indigenous language. Setting the default language of the Niupepa digital library to reo Māori not only resulted in significantly more requests with the interface language set to reo Māori but also resulted in te reo Māori being used in more situations to undertake information retrieval. Examples of these situations included a higher use of the series list (where te reo Māori is prominent), a higher number of reo Māori documents accessed, longer read times of those documents and a higher percentage of query terms written in te reo Māori.

This chapter has shown that the assistance of a bilingual digital library to the usage of an indigenous language can increase in two areas if the default language of the digital library is set to the indigenous language. First, the amount of times that the indigenous language is displayed and, subsequently, used by clients is significantly increased. Second, the new opportunities and avenues available for using the indigenous language are exploited more often if the default language is set to the indigenous language.

6 Hawaiian Language Usage of the Nūpepa Digital Library

Chapter 6 discusses how the Hawaiian language was used in 2005 by clients of the Hawaiian Nūpepa digital library—Nā Nūpepa ‘Ōlelo Hawai‘i. Hereafter, this digital library will be referred to simply as the Nūpepa digital library and should not be confused with the Niupepa digital library which refers to the Aotearoa Māori Niupepa digital library.

The Nūpepa digital library was another digital library suitable for analysing the use of an indigenous language in a modern electronic medium. Like its Māori counterpart, it provided a significant quantity of indigenous language resources, through an indigenous and non-indigenous switchable interface and was freely available on the Internet.

Was the Hawaiian language used by clients of the Nūpepa digital library to access documents written in Hawaiian? Results from TLA undertaken on the Nūpepa dataset indicate that the Hawaiian language was the preferred language in over 200,000 requests to the Nūpepa digital library in 2005. Given that the Hawaiian language was used by Nūpepa clients, the subsequent question to pose is how was it used?

This chapter details how the Hawaiian language was used in the Nūpepa digital library in 2005. The chapter begins with an introduction of the Hawaiian Nūpepa digital library, where the content was sourced, how the digital library was established and what differences there were between the Hawaiian Nūpepa digital library interface and the Māori Niupepa digital library interface. This is followed

by a section which details how much the Hawaiian language was used in the Nūpepa digital library in 2005.

Exploratory session data is analysed to determine how Hawaiian language clients explored the Nūpepa digital library but then did not go on to use it. This data reveals that Hawaiian exploratory sessions on average lasted less than three minutes and consisted of three requests or fewer.

Usage session data will be analysed in three sections; how Hawaiian language clients used the Nūpepa interface, how they accessed Nūpepa documents and submitted queries. The three sections provide an extensive examination of how the Hawaiian language was used by Nūpepa clients.

Two further topics are examined using statistics generated from TLA. First, Hawaiian language usage of the Nūpepa digital library is compared to English language usage. This will reveal that Hawaiian language clients undertook longer sessions, used different methods to access documents and preferred different document formats from English language clients. The second and final topic examined is the use of the Nūpepa digital library by clients who preferred to access documents bilingually, that is, using both Hawaiian and English. A quarter of all requests were recorded in sessions where clients used the Nūpepa interface bilingually. Bilingual clients exhibited unique querying behaviours that led to a high number of documents accessed per query.

Chapter 6 concludes by summarising how the Hawaiian language was used by clients of the Nūpepa digital library.

6.1 The Hawaiian Nūpepa Digital Library

The next four sub-sections give introductory information about the Hawaiian Nūpepa digital library. A brief history is given about the content of the Nūpepa digital library, newspapers written in Hawaiian, and how these were captured and placed in an online environment. This is followed by a description of the Nūpepa interface, which includes screenshots taken from the Nūpepa web site. This section concludes by detailing the differences between the Nūpepa and Niupepa interfaces.

6.1.1 Capturing the Hawaiian Nūpepa

The publication of newspapers in Hawai‘i began early in its colonial history. The first two newspapers, *Ka Lama* (The Torch) and *Ke Kumu* (The Teacher), both published in 1834, were Protestant mission-sponsored newspapers. These publications began a myriad of newspaper publications that occurred in the Hawaiian Islands. Some estimates suggest over 1,000 separately-titled newspapers have been published there in at least nine different languages (Chapin, 1996), of which just over 100 were in the Hawaiian language. Although the newspapers were diverse, Chapin categorises them into four main groupings: establishment, opposition, official and independence. The amount of information and the unique viewpoints ensure that the newspapers are rich and valuable sources of historical information.

In the mid-1980s, efforts began in the capture and preservation of the newspaper pages that had survived until then. Although the large majority of newspaper publications had not been preserved, there were still significant numbers of newspaper pages that had been stored in the various libraries, archival collections and museums throughout Hawai‘i. A grant was awarded to the University of Hawai‘i at Manoa, by the National Endowment for the Humanities, to copy all remaining historic Hawaiian and ethnic newspapers onto microfilm. In total 260,000 pages were microfilmed, including papers in Hawaiian, English and

Portuguese. While there were names for 103 known Hawaiian language newspapers, only about half of these had surviving hard copies that were available to be microfilmed. These hard copies comprised about 60,000 discrete newspaper pages.

The Hawaiian newspapers are written predominantly in the Hawaiian language. It was a surprise to researchers recovering the images to discover that some of the pages were, in fact, written in English. These writings are small in number, sometimes a column, sometimes a page, and are scattered throughout some of the newspaper pages in a small number of the periodicals. Although there are no exact figures it appears the English writings comprise an insignificant proportion of the writings in the Hawaiian language newspapers.

6.1.2 Hawaiian Nūpepa Online

Work on placing the Hawaiian language newspapers online began in 1997. Some of the microfilm copies of the newspapers had begun to deteriorate as a result of frequent use by students and scholars of the Hawaiian language. Three libraries of the University of Hawai'i received a small grant to undertake a pilot project to digitise a small number of the heavily used newspapers pages and place them on the web. Further small grants were received in 1998, 1999 and a more substantial grant in 2000, which allowed more titles and more newspaper pages to be added to the online collection (Cartwright, Chantiny, Hori, & Peacock, 2000). At the completion of this funding approximately 2,000 pages of Hawaiian language newspapers could be viewed online.

Having the Hawaiian language newspapers online created much interest throughout the Hawaiian language communities, and with researchers and scholars of Hawaiian language and history. Investigations began into the possibility of placing all Hawaiian language newspaper pages online and making them full text searchable. Initial discussions involved Dwayne Nakila Steele, a philanthropist and retired businessman, Puakea Noglemeier, professor of Hawaiian Language at the University of Hawai'i at Manoa, the Hawaiian

Language Legacy Programme of Alu Like Inc.ⁱ, and Hale Kuamo‘oⁱⁱ. Alu Like Inc. had at that time been working with Hale Kuamo‘o to place a journal of Hawaiian literature online, so had already established a working relationship and had begun looking at the issues of hosting a digital library web site. Steele provided personal funding and was able to secure funding from other private sources.

Capturing the text from the Hawaiian newspaper pages proved to be difficult for three reasons. First, the volume of the text per page was very large; tabloid (597mm × 375mm) and broadsheet (749mm × 597mm) page sizes were common in the newspapers many of which can consist of up to six columns. On average there were over 6,300 words per newspaper page. Second, the quality of the captured image was often poor; smudging, fading, skewed text or gutter/baseline defects were often present in the newspaper pages, all of which reduce the ability of the OCR process to accurately generate texts. Third, there was very little computer software support for the Hawaiian language; consequently the OCR program had to be reconfigured and dictionaries had to be created to enable the OCR process to read the Hawaiian language.

The extracted text was generated automatically and, in the early stages of the web site development, there was minimal work done on correcting the errors resulting from this process. However, after much debate it was decided that the error rate was unacceptable and that Hawaiian language experts were needed to carefully correct the texts generated from the OCR process. In 2003, the Ho‘olaupa‘i

ⁱ Alu Like Inc. is the largest non-government, non-profit service organisation that assists native Hawaiians to achieve social and economic self sufficiency. Alu Like Inc. has services which include economic and business development, employment preparation and training, library services, education and childcare services. Alu Like Inc. is funded by federal, state, county and private sources and is governed by a state-wide volunteer Board of Directors.

ⁱⁱ Hale Kuamo‘o is the the Hawaiian Language Center within Ka Haka ‘Ula O Ke‘elikōlani (the Hawaiian Language College) of the University of Hawai‘i at Hilo. Hale Kuamo‘o was established by Hawai‘i State Legislature in 1989 to encourage, expand and sustain Hawaiian language, Hawaiian culture and Hawaiian-medium education in all situations. Its activities include providing publications and training for Hawaiian-medium teachers, making available Hawaiian literature for various forms of media and undertaking research into Hawaiian dictionaries and grammar terminology.

project was established at the Bishop Museum in Honolulu to partner with Alu Like Inc. and Hale Kuamo‘o and undertake OCR of the Hawaiian Newspapers to a high standard of accuracy (upwards of 99.8%). Hawaiian language students were employed to undertake the laborious task of OCR and Hawaiian language experts were used to verify the extracted texts. However, because of funding constraints, the difficulties listed above and the level of accuracy required, progress on getting all the Hawaiian language pages converted into an accurate text has not been rapid. By early 2007 approximately 100 to 150 extracted text pages were being generated per month.

The newspapers are located on a web site called Ulukau—the Hawaiian Electronic Library web site. As mentioned in §2.2 Ulukau is also the name of the web server that makes available the Nūpepa collection. The Ulukau web site can be found online at <http://ulukau.org>. A private company called DL Consulting Ltd, located in Hamilton, Aotearoa, was contracted to manage the process of making these and other Hawaiian language resources available via the Internet. DL Consulting built the Ulukau web site using the open source digital library software called Greenstone (Greenstone software is described in §1.2.3). The purpose of the Ulukau web site is to make Hawaiian language resources available for the use, teaching and enhancement of the Hawaiian language.

The Ulukau web site was established in 2004 and operates through the co-operation of dozens of agencies and individuals. Its leading two partners are Hale Kuamo‘o and the Native Hawaiian Library. The Native Hawaiian Library is one of the major programmes funded by Alu Like Inc., Inc. It provides library and information services to native Hawaiians and others who are interested in Hawaiian information and documents. The Ulukau web site is an ongoing web site with more documents and more collections being added as they become available.

As at June 2006 there were seven electronic collections publicly available at this web site:

- Puke Pai ‘Ōlelo Hawai‘i—Journal Of Hawaiian Language Sources
- Nā Palapala o ka Māhele—Māhele (land) Database
- Nā Nūpepa ‘Ōlelo Hawai‘i —Hawaiian Newspapers
- Nā Inoa ‘Āina—Hawaiian Place Names
- Ka ‘Ohina ki‘i Na Ed Greevy—Ed Greevy’s Photograph Collection
- Nā Puke Wehewehe ‘Ōlelo Hawai‘i —Hawaiian Dictionaries
- Nā Puke ‘Ē A‘e—More Books

6.1.3 Hawaiian Nūpepa Interface Description

The Hawaiian language newspapers, located on the Ulukau web site can be accessed directly via the URL www.nupepa.org. By July 2006 there were 53,425 Hawaiian language newspaper pages available on this web site. According to the Hawaiian Language Legacy Programme (Stauffer, 2006), these pages constitute nearly all of the surviving Hawaiian language newspaper pages. These are from just over 100 reels of microfilm and represent 46 different newspapers. The fewest number of surviving pages from a newspaper was two pages from *Ko Hawai‘i Pae Aina*, while the largest number of surviving pages was 18,153 pages from the periodical *Ka Nupepa Kuokoa*.

No exact records were kept for the year of 2005; however, a calculated estimate would suggest that in 2005 there were an average of 2,550 text files available out of a total of 40,000 discrete documents, representing 6.4% of the total documents.

The Nūpepa web site is a bilingual web site that, by default is displayed in the Hawaiian language. There is a button located in the top right-hand corner that can alternate the interface language display between English and Hawaiian. This button is available on all pages of the Nūpepa web site. The Nūpepa home page gives some explanatory information about the Nūpepa Collection and makes available three methods to access the newspapers; search (*huli*), titles (*inoa nūpepa*) and dates (*lā ho‘opuka*). The Nūpepa web site home page is displayed in *Figure 6-1*.

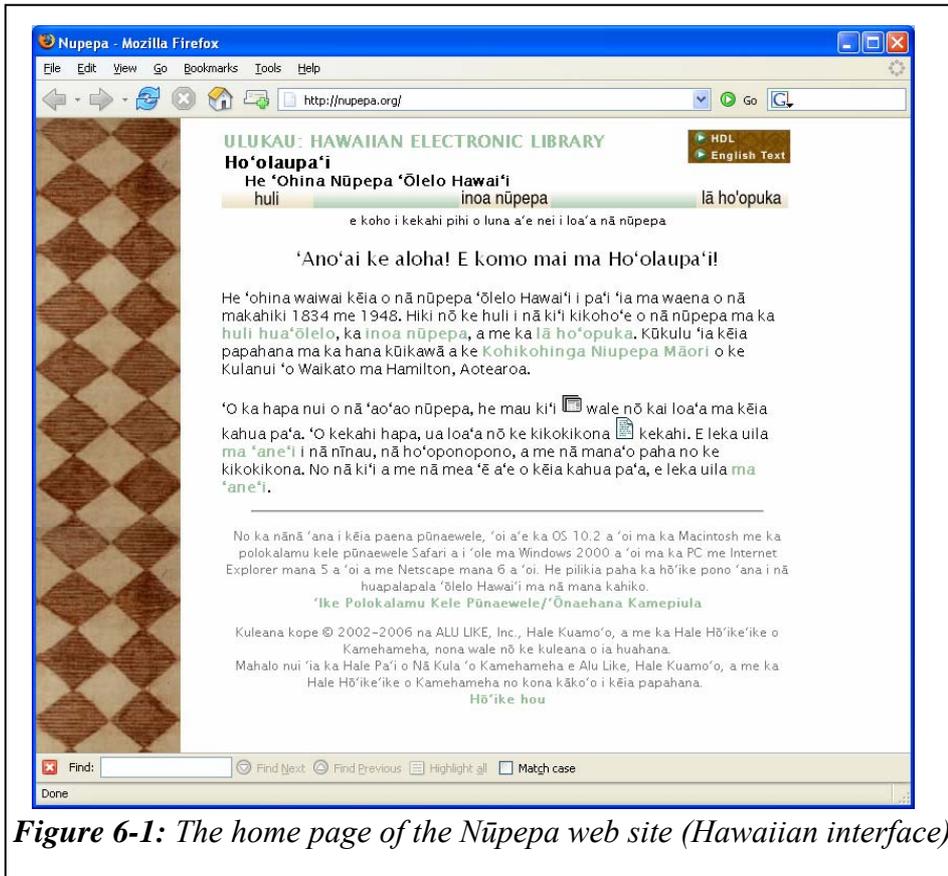


Figure 6-1: The home page of the Nūpepa web site (Hawaiian interface).

The first method available to access Nūpepa documents is by undertaking a query (huli). The English version of the query page is displayed in *Figure 6-2* and shows the results of a query for the term *'nupepa'*. It can be seen that the query interface screen is simple, with just the one query argument able to be altered; the Some/All option. Clicking on an icon underneath the results bar results in either a graphic page or an extracted text document being displayed (as in *Figure 6-3*).

When the extracted text page is displayed, the query term is highlighted as can be seen in *Figure 6-3*. This page provides links to access the previous newspaper page, the next newspaper page, or any page in the current edition of the newspaper. There are also links to a preview image, a mid-size image, a full size image, a link to detach the text to a separate browser window and a link that will display the text without any highlighting.

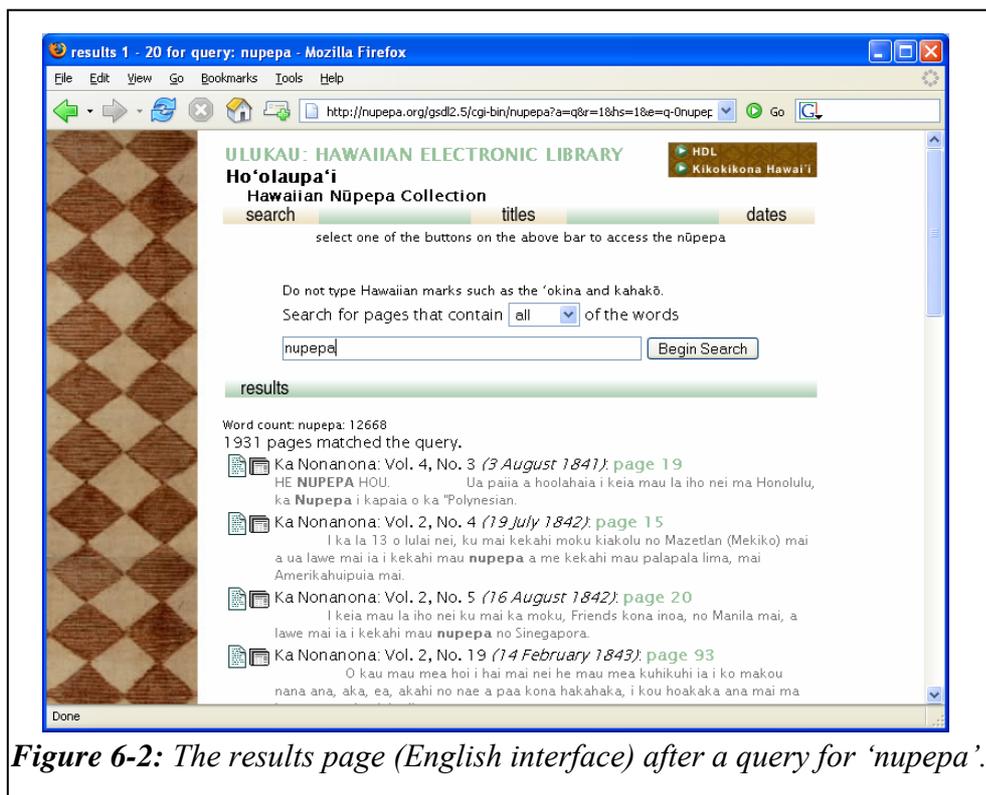


Figure 6-2: The results page (English interface) after a query for 'nupepa'.

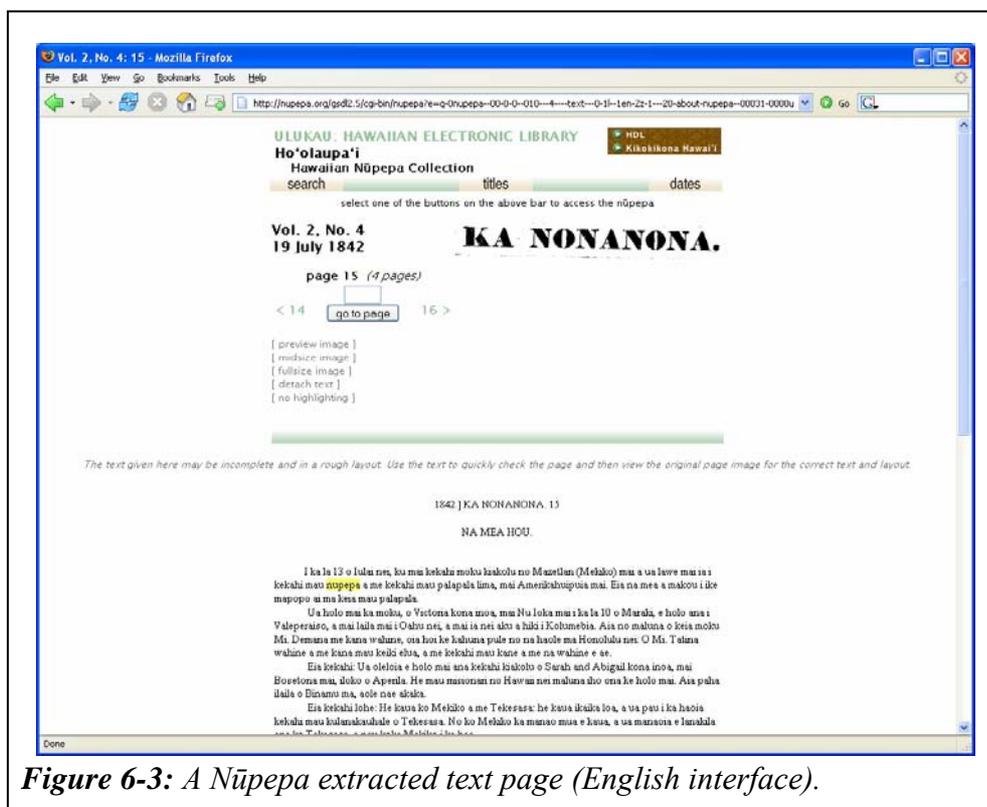


Figure 6-3: A Nūpepa extracted text page (English interface).

The second method of accessing Nūpepa documents is to select a document from the series list. Clicking on the titles (*inoa nūpepa*) button will display a series list of titles as shown in *Figure 6-4*.

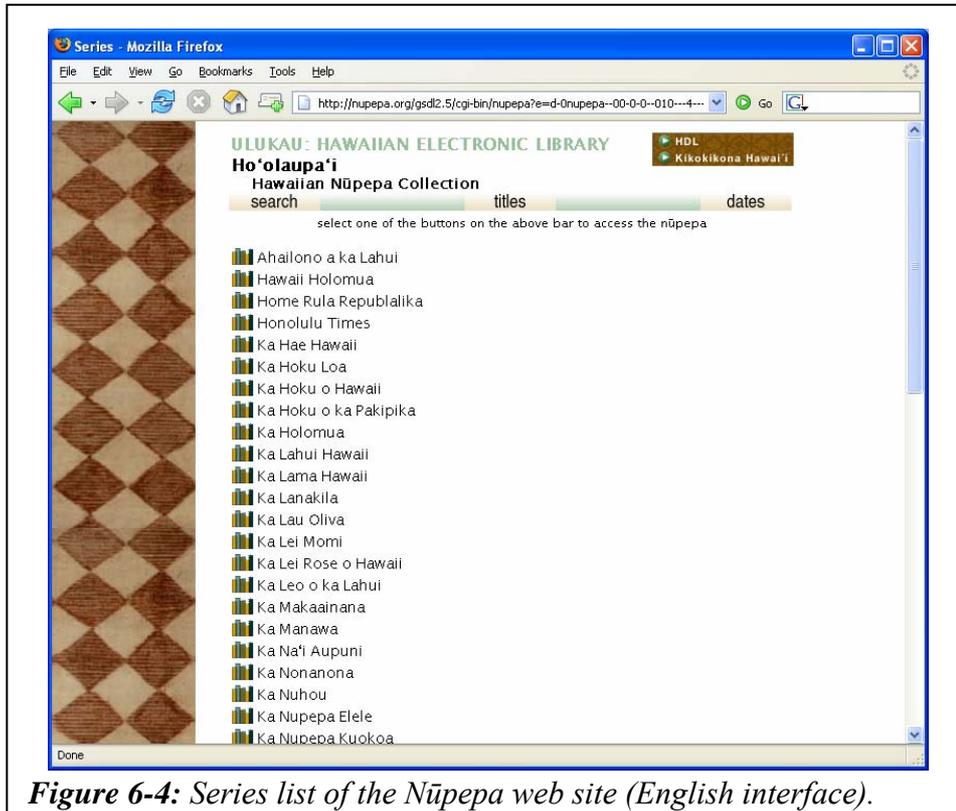


Figure 6-4: Series list of the Nūpepa web site (English interface).

The third and final method of accessing Nūpepa documents is to select a document from a grouping by date. Clicking on the dates (*lā ho'opuka*) button and then the year 1862 will lead to the screen displayed in *Figure 6-5*.

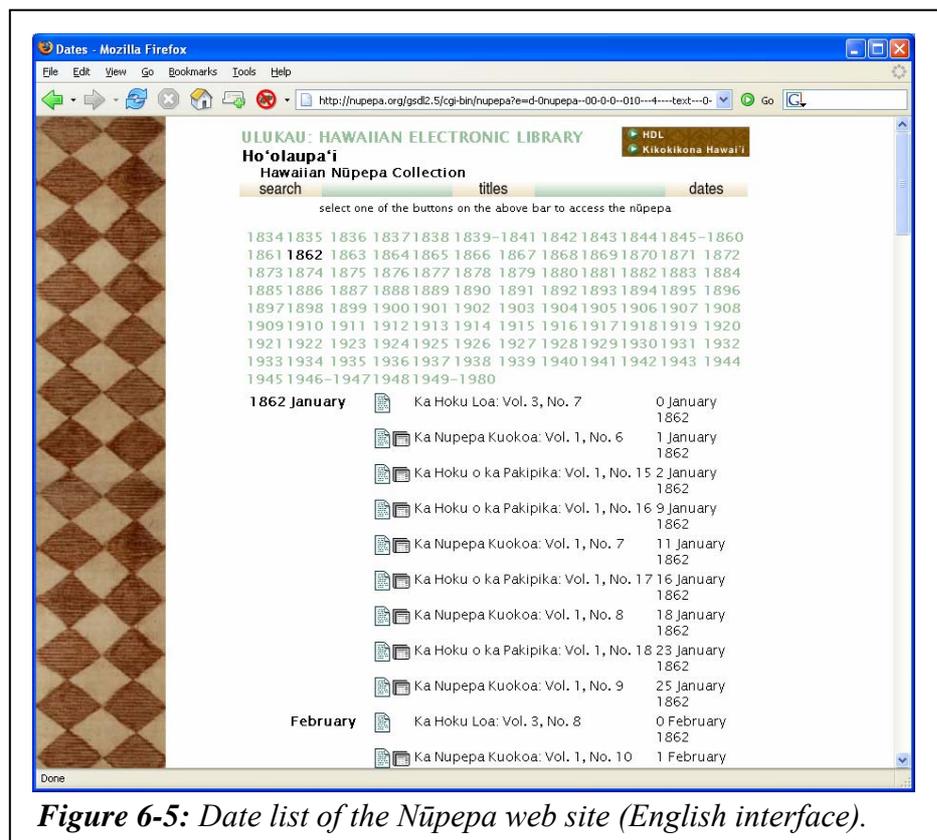


Figure 6-5: Date list of the Nūpepa web site (English interface).

6.1.4 Nūpepa/Niupepa Interface Differences

While the Hawaiian Nūpepa digital library shares some significant similarities with the Māori Niupepa digital library there are some important interface differences. These interface differences affect how Hawaiian Nūpepa clients and Māori Niupepa clients interact with their respective digital libraries and should be kept in mind when usage of these interfaces is compared. The differences are summarised in *Table 6-1*:

The first item listed in *Table 6-1* is the disabling of browser caching. The Ulukau web server was set up to disable browser caching. The HTTP headers that are sent with the web pages have a 'must revalidate' command written in them that instructs the users' browsers not to cache these pages. This feature was not used in the Māori Niupepa web site. It is difficult to determine how effective this command is at disabling browser and network caching and the subsequent affect it has on web log statistics.

Interface Features of Nūpepa and Niupepa Digital Libraries		
	Hawaiian Nūpepa	Māori Niupepa
Disable browser caching	Yes	No
Preferences page	No	Yes
Help page	No	Yes
Interface languages available	2	43
Language switch available on all pages	Yes	No
Full size images	~40,000	35,482
Mid-size images	~40,000	0
Preview image	~40,000	35,482
Extracted text generated from images	~6.4%	98.6%
Detach text	Yes	No
Parallel (indigenous/non-indig.) documents	<1%	44.2%
English language documents	0%	1.9%
Abstracts in English	0%	43.5%
Commentary information	0%	64.7%

Table 6-1: Comparison of Nūpepa and Niupepa interface features as available to clients in 2005.

As shown in *Table 6-1* some features available in the Māori Niupepa digital library were not available in the Hawaiian Nūpepa digital library. The features not available in the Hawaiian Nūpepa digital library were: a preferences page, a help page, a multilingual interface, a high proportion of extracted text documents, a significant number of parallel (indigenous language/non-indigenous language) documents, Abstracts in English and commentary information.

The Hawaiian Nūpepa digital library had some features that were not available in the Māori Niupepa digital library. These extra features included: the disabling of browser caching, the ability to switch the language of the interface (Hawaiian to English and English to Hawaiian) on every interface page, the option to detach (open) the text in another browser window, mid-size images and the ability to view full size images in a pdf format. The pdf format gave the client the ability to zoom and pan on the image, and allowed easier printing of the image.

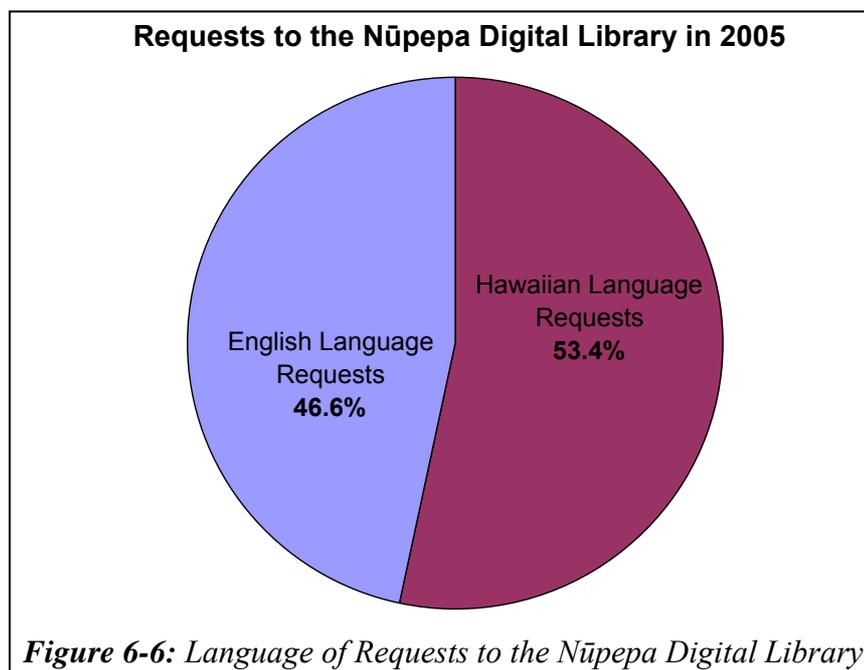
Two further features of the Hawaiian Nūpepa digital library meant it was more likely than the Māori Niupepa digital library to encourage the use of the indigenous language. First, virtually all the content was available only in the Hawaiian language. This meant that the resource was of prime benefit to Hawaiian-speaking clients and of little benefit to clients who could not understand Hawaiian. In contrast, the Māori Niupepa digital library included a significant proportion of English-language material, which benefited clients who only spoke English. The material comprised 44% parallel language (reo Māori/English) Niupepa documents, 2 % newspaper documents published in English, and 44.5% of reo Māori documents that had additional summary abstracts written in English.

The default language setting was the second feature of the Hawaiian Nūpepa digital library that encouraged use of the indigenous language. The Nūpepa default interface language was set to Hawaiian for the whole of 2005. The previous chapter showed that an indigenous language setting not only increases the number of requests recorded in the indigenous language but also increases indigenous language use of other aspects of the interface.

6.2 Was the Hawaiian Language Used?

Transaction Log Analysis (TLA) of the Hawaiian Nūpepa digital library was used to answer whether the Hawaiian language was used by Nūpepa clients.

The Hawaiian Nūpepa web site received a total of 286,383 raw requests in 2005. The first request was logged at 01:17am January 1st 2005 and the final request was logged at 11:59pm December 31st 2005. This raw log of requests was then prepared by removing all redundant CGI arguments and extraneous requests (using the method described in §2.3), leaving a filtered total of 211,215 requests. Of this total, 112,786 were logged as requests with the interface language set to Hawaiian and 98,429 were requests with the interface language set to English. These requests are displayed as percentages in *Figure 6-6*.

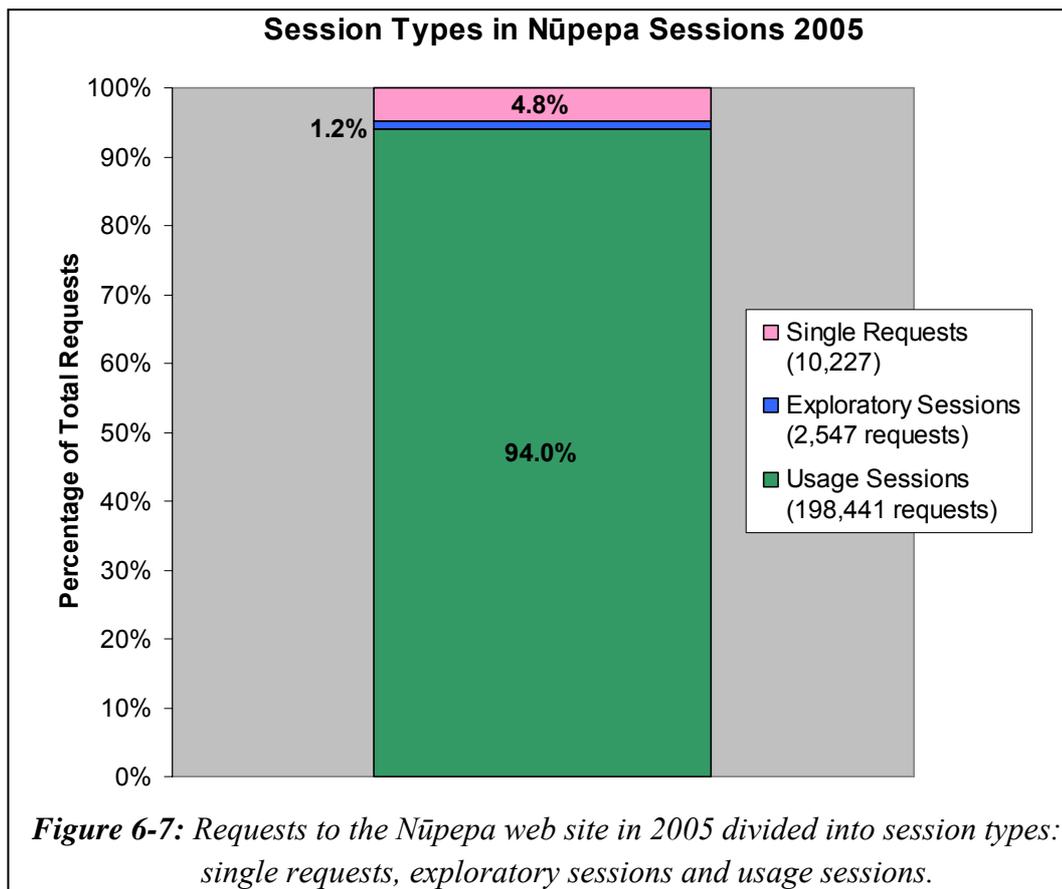


Clearly the Hawaiian language was used by Nūpepa clients; the interface was set to Hawaiian in more than half of the requests made to the Nūpepa digital library.

6.2.1 Was Hawaiian Used in All Three Session Types?

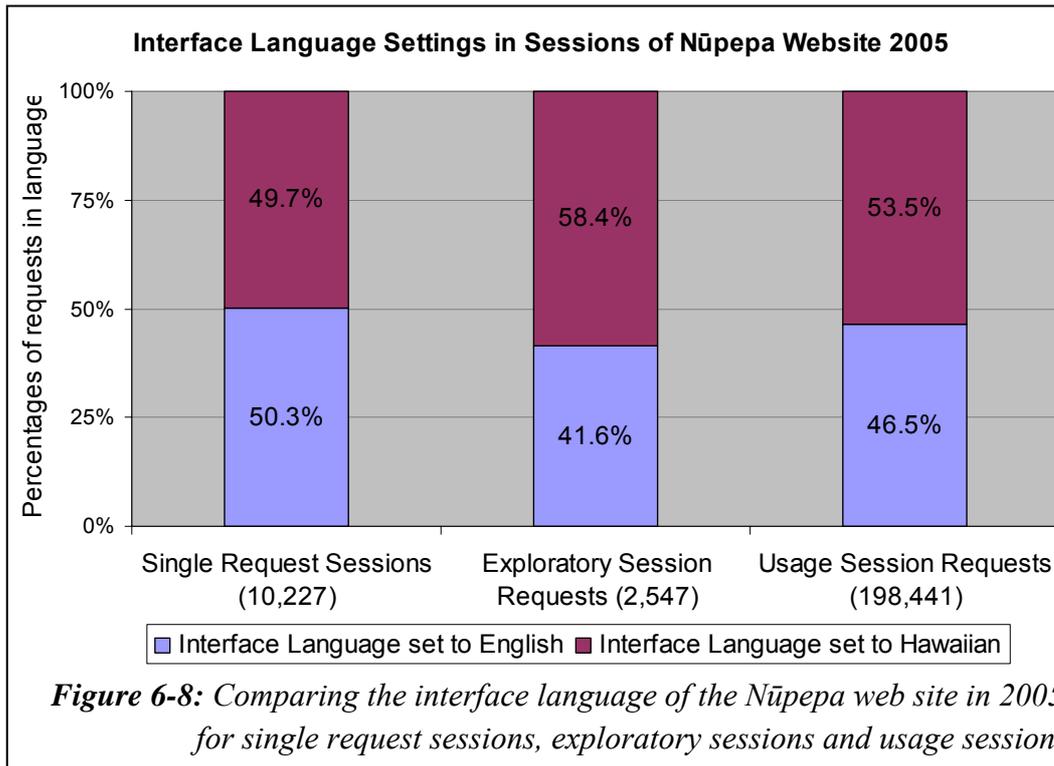
The TLA process collated the requests into three types of sessions. There were 10,227 requests (4.8%) classified as single requests, 2,547 requests (1.2%) classified into 948 exploratory sessions and 198,441 requests (94.0%) classified into 8,754 usage sessions. The percentages of requests per session type are displayed in *Figure 6-7*.

Figure 6-7 shows the majority of the requests (94.0%) to the Nūpepa digital library were classified as usage session requests; requests from clients who were actively engaging with the Nūpepa digital library to retrieve information.



The three session types were further examined to determine which interface language was preferred by Nūpepa clients in 2005. For single requests, 5,085 requests had the interface language set to Hawaiian and 5,142 had the interface

language set to English. For exploratory sessions, 1488 requests had the interface language was set to Hawaiian and 1059 had the interface language was set to English. For usage sessions, 106,213 requests had the interface language set to Hawaiian and 92,228 had the interface language set to English. These figures have been calculated as percentages and are graphed in *Figure 6-8*.



The Hawaiian language was the preferred language of use. *Figure 6-8* clearly shows higher percentages of requests in Hawaiian than in English.

The fact that 53.5% of requests in usage sessions are in Hawaiian is noteworthy as it suggests two important findings. First, not only are Nūpepa clients willing to undertake exploratory and browsing activity in Hawaiian; they are also willing to undertake information retrieval in Hawaiian. Second, it suggests that setting the default interface language to Hawaiian is appropriate for most of the usage activity undertaken in the Nūpepa digital library.

6.2.2 Was the Hawaiian DL Used Bilingually?

Clients could use the Nūpepa digital library in Hawaiian, in English or in both languages, that is, bilingually. A bilingual Nūpepa session was defined the same as a bilingual Niupepa (Māori) session (see §2.4.3). There were 1,090 bilingual sessions in 2005, resulting in 50,359 usgBi requests to the Nūpepa digital library. Bilingual sessions comprised 13,541 (26.9%) requests using Hawaiian and 36,818 (73.1%) requests using English. Bilingual sessions made up 12.5% of the total usage sessions and 25.4% of the total usage requests to the Nūpepa digital library.

The Hawaiian Nūpepa digital library was used bilingually. One in eight usage *sessions* and one in four usage *requests* were categorised as bilingual.

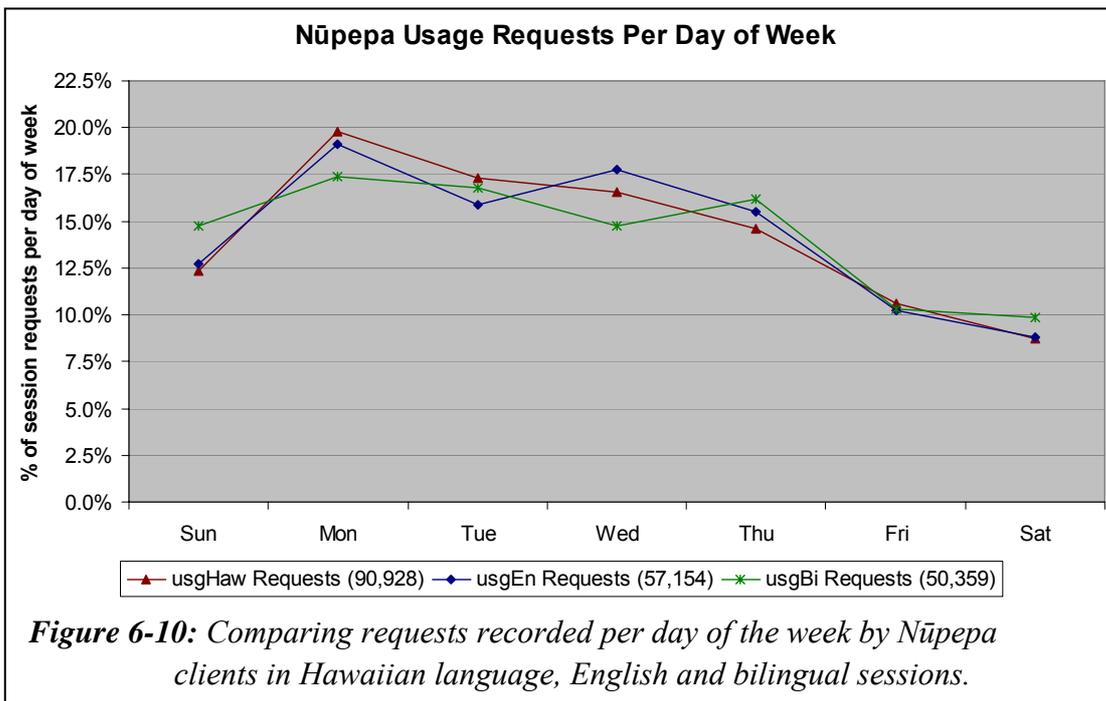
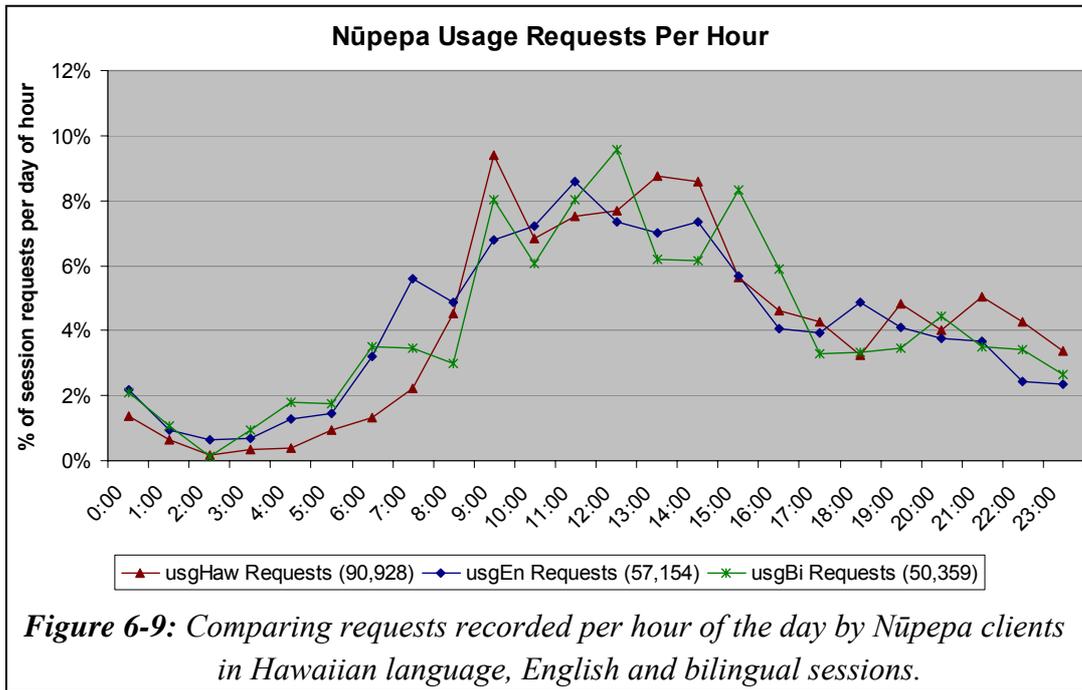
6.2.3 When Was the Hawaiian Language Used?

The times that usage requests were received (time of day, day of week, and month of year) are shown for each language type in *Figures 6-9, 6-10 and 6-11*.ⁱⁱⁱ The sessions have been plotted as percentages of requests per session to take into account the variation in frequency of requests in the different session types.

Figure 6-9 shows the requests as they occurred in a 24 hour time period. The requests occurred at similar times no matter which interface language was being used by the Nūpepa clients. The 9 hours between 8:00am and 5:00pm, the standard work hours, was when most of the activity occurred; 62% of all requests were logged in this time period.

Figure 6-10 shows the requests as they occurred each day. Most of the activity occurred Monday to Thursday, with the least activity occurring on a Saturday.

ⁱⁱⁱ Note: no allowance was made in these graphs for the 6.7% of clients who were making requests from a different time/date zone.



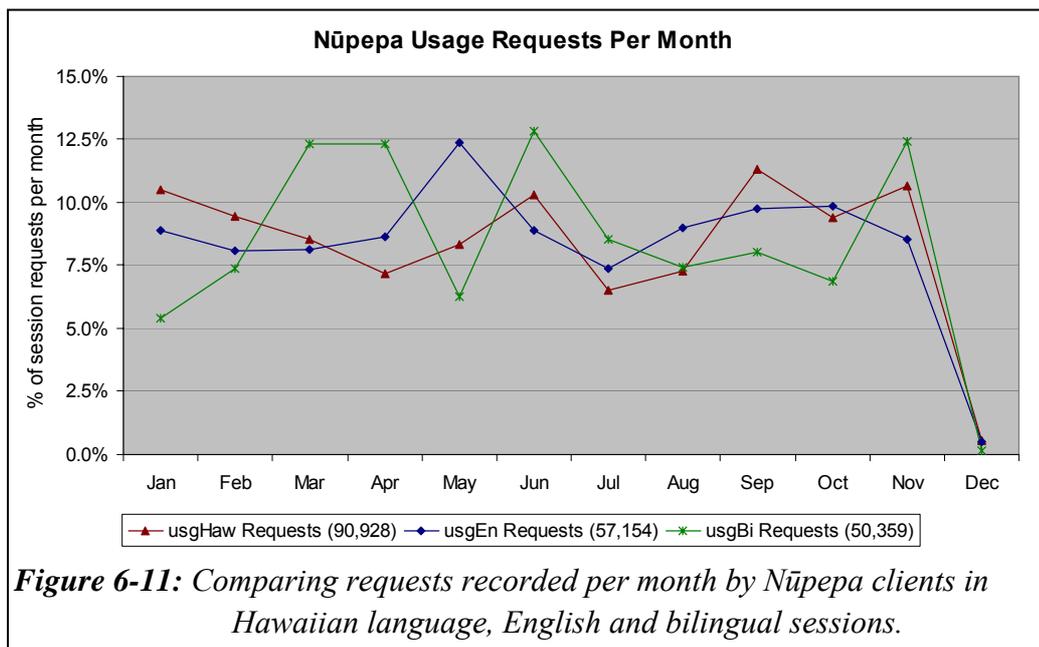


Figure 6-11 shows the requests as they occurred throughout 2005, grouped by month. The possibility of some data being compromised because of a configuration setting of the Ulukau web server, as described in §2.2.4, led to the removal of almost all the December requests from this analysis.

A similar proportion of Hawaiian (usgHaw) and English language usage sessions (usgEn) occurred per month. While the proportion of bilingual usage sessions (usgBi) per month was not as closely aligned, the difference was never more than 6.2%. The Nūpepa interface language was consistently used in Hawaiian, English and bilingually throughout 2005.

6.2.4 Summary—Was the Hawaiian Language Used?

The Hawaiian language was used in more than half of the requests submitted by Nūpepa clients. Approximately half of the Nūpepa usage requests were recorded as usgHaw sessions, one quarter as usgEn sessions and one quarter as usgBi sessions. UsgHaw, usgEn and usgBi clients used the Nūpepa digital library in similar chronological periods; mostly 9:00am to 5:00pm, Monday to Thursday, in all months of the year.

6.3 Exploratory Behaviours

Requests made by clients who explored the Nūpepa digital library but then did not go on to use it represent only 1.2% of the total requests. In total, 2547 requests were defined into 948 Nūpepa exploratory sessions. Nūpepa exploratory sessions were usually short sessions; a high proportion (84.0%) was less than three minutes and (93.7%) consisted of three requests or less. Two requests were the minimum requirement for classification as a session.

Nūpepa exploratory sessions were subdivided into two language categories. There were 1,589 requests in 537 exploratory sessions where the interface language was mostly set to Hawaiian (expHaw). In these sessions the interface language was set to Hawaiian in 72.6% of the requests. There were 958 requests in 411 exploratory sessions where the interface language was mostly set to English (expEn). The interface language was set to English in 64.6% of these requests.

What information can be gained from the TLA process to illustrate how the Nūpepa digital library was explored but then not used by Hawaiian language clients? The Hawaiian Nūpepa interface was simpler than the Māori Niupepa interface as there was no help page or preferences page. This meant that an analysis of which interface pages were requested could not be undertaken. Only two analyses could be made, the length of exploratory sessions and the apparent read times in exploratory sessions.

6.3.1 How Long Were Exploratory Sessions?

Session length is measured in two categories, duration in time (seconds) and length in number of requests. The comparative expHaw and expEn lengths of sessions in seconds are displayed in *Table 6-2*. The comparative expHaw and expEn lengths of sessions in number of requests are displayed in *Table 6-3*.

expHaw			expEn		
Seconds	Sessions	%	Seconds	Sessions	%
Total	537	100%	Total	411	100%
0–1	51	9.5%	0–1	10	2.4%
2–9	49	9.1%	2–9	146	35.5%
10–29	205	38.2%	10–29	133	32.4%
30–179	163	30.4%	30–179	80	19.5%
180+	69	12.8%	180+	42	10.2%

Table 6-2: Comparisons of lengths (in seconds) of expHaw and expEn sessions.

expHaw			expEn		
Requests	Sessions	%	Requests	Sessions	%
Total	537	100%	Total	411	100%
2	177	33.0%	2	343	83.5%
3	270	50.3%	3	41	10.0%
4	51	9.5%	4	15	3.6%
5+	39	7.3%	5+	12	2.9%

Table 6-3: Comparisons of number of requests in expHaw and expEn sessions.

ExpHaw sessions were short; 56.8% were less than 30 seconds and 87.2% were less than 3 minutes. In terms of number of requests 83.3% consisted of just two or three requests.

Both tables show that expHaw sessions were slightly longer than expEn sessions. *Table 6-2* indicates that expHaw sessions had a higher proportion of 30–179 second sessions (30.4% compared with 19.5%). The average duration of expHaw sessions was 1.63 minutes while the average duration of expEn sessions was 1.56 minutes. *Table 6-3* indicates that expHaw sessions had a much higher proportion of sessions involving three requests (50.3% compared with 10.0%). The average number of requests in expHaw sessions was 3.0 requests compared with 2.3

requests in expEn sessions. While these differences may not appear large, with p values of $3.9 * 10^{-23}$ and $2.0 * 10^{-52}$ respectively, the differences were statistically significant.

While both expHaw and expEn sessions were short, the expEn sessions were shorter than the expHaw sessions. The difference between the length of expHaw sessions and the length expEn sessions was statistically significant.

6.3.2 How Long to Read Interface Pages?

The lengths of times between requests for pages to the Nūpepa web site were calculated and recorded as apparent page read times. These read times were grouped into time intervals of, 1–2 seconds, 3–9 seconds, 10–19 seconds, 20–39 seconds, 40–79 seconds and 80+ seconds. These times are shown in *Table 6-4*.

The average apparent read time of interface pages by expHaw clients was 61.4 seconds. The average apparent read time of interface pages by expEn clients was 83.8 seconds. Chi square analysis on the read times listed in *Table 6-4* returned a p value of 0.22. This p value indicates that there was no statistically significant difference in the expHaw and expEn apparent read times of Nūpepa interface pages.

Seconds	expHaw	expEn
Total Apparent Reads	1052	547
1–2	103	37
3–9	362	195
10–19	257	132
20–39	145	79
40–79	56	41
80+	129	63

Table 6-4: expHaw and expEn apparent read times of Nūpepa interface pages in exploratory sessions.

6.3.3 Summary—Exploratory Behaviours

Exploratory sessions in Hawaiian, like exploratory sessions in te reo Māori, were short. ExpHaw sessions mostly consisted of two or three requests and half of the sessions were completed within 30 seconds. Almost half of the apparent read times of interface pages by expHaw clients were less than 10 seconds.

While exploratory activity by Hawaiian language clients was short, it wasn't as short as exploratory activity by English language clients. ExpEn clients undertook shorter exploratory sessions, both in duration and number of requests, than expHaw clients. The apparent read times of interface pages by expHaw clients and expEn clients were statistically similar.

Exploratory clients only request pages from the interface; no documents are requested and no queries are submitted. Aside from session lengths and apparent read times, it is difficult to derive information about how the Hawaiian language was used by exploratory clients of the Nūpepa digital library. Exploratory session requests comprised just a small (1.2%) portion of the activity that occurred in the Nūpepa digital library in 2005.

The next three sections analyse how the Hawaiian language was used by clients to actively engage with the Nūpepa digital library to retrieve information. Usage of the Nūpepa digital library by usgHaw clients is analysed in three parts; interface behaviours, document access behaviours and query behaviours. The first part, an analysis of how usgHaw clients used the Nūpepa interface, begins in the following section.

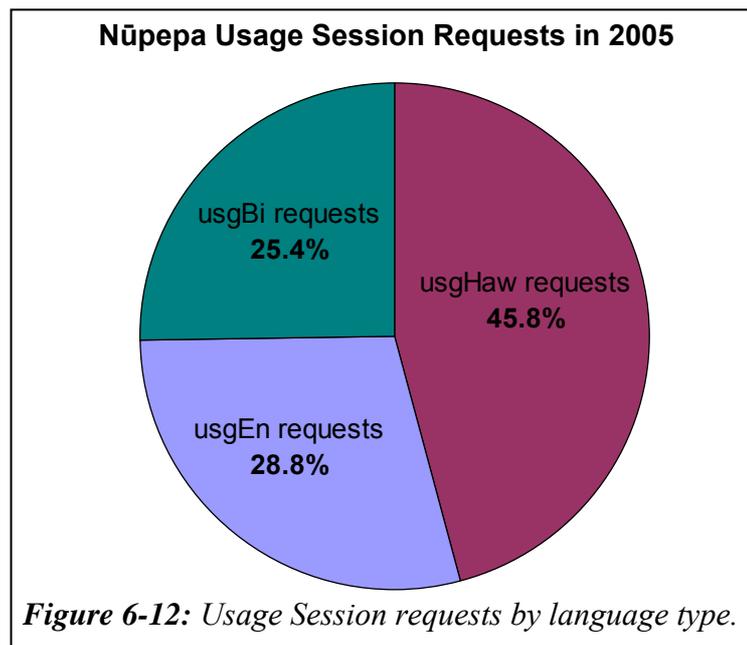
6.4 Nūpepa Interface Behaviours

This section analyses how the Nūpepa interface was used by Hawaiian language clients in usage sessions. A usage session is a session where the client actively engages with the Nūpepa interface to access Nūpepa documents. Data about usage session activity is the most important session data to analyse as it represents the majority (94%) of activity and specifically targets the key target group, clients who actively use the Nūpepa digital library.

There were 198,441 requests recorded in 8,754 usage sessions in the Hawaiian Nūpepa digital library in 2005. Of the requests recorded, 106,213 (53.5%) used the interface in Hawaiian and 92,222 (46.5%) used the interface in English. On average, the usage sessions undertaken by Nūpepa clients were 10 times longer in duration and eight times longer in number of requests than the exploratory sessions.

The Nūpepa usage session requests were then subdivided into language type (as described in §2.4.3). There were 90,928 requests in 4,182 usage sessions when the language of the interface was mostly set to Hawaiian (usgHaw) and 57,154 requests in 3,482 usage sessions when the language of the interface was mostly set to English (usgEn). There were 50,359 requests in 1,090 usage sessions when the language of the interface was switched between Hawaiian and English; these sessions were subsequently defined as bilingual sessions (usgBi). An analysis of the usgHaw requests shows the interface language set to Hawaiian in 98.6% of the requests. An analysis of the usgEn requests shows the interface language set to English in 94.7% of the requests. An analysis of the usgBi requests show the interface set to English in 73.1% of the requests and Hawaiian in 26.9% of the requests.

The usage session requests for each language type were calculated as percentages and are displayed in *Figure 6-12*.



Statistics generated from TLA were used to examine six interface behaviours exhibited by Nūpepa clients who used the Hawaiian language:

- the length of usgHaw sessions
- the apparent read times of interface pages in usgHaw sessions
- the number of return usgHaw sessions
- the origin of usgHaw sessions
- the frequency of interface language switching
- the position of interface language switching in usgHaw sessions.

These six interface behaviours comprise the next six sub-sections.

6.4.1 How Long Were Usage Sessions?

Session length is defined by two measurements; duration in time (seconds and minutes) and length in number of requests. Statistics generated from TLA on the Nūpepa digital library are listed to show the length of usgHaw sessions in 2005. These statistics are listed alongside length statistics on usgEn and usgBi sessions.

How Long in Time?

The durations of usgHaw sessions were grouped into the following time intervals; 0–10 seconds, 11–60 seconds and 1–3 minutes, 3–15 minutes, 15–30 minutes, 30+ minutes. The numbers of sessions that fitted into these categories were calculated as percentages for comparison purposes and graphed in *Figure 6-13*.

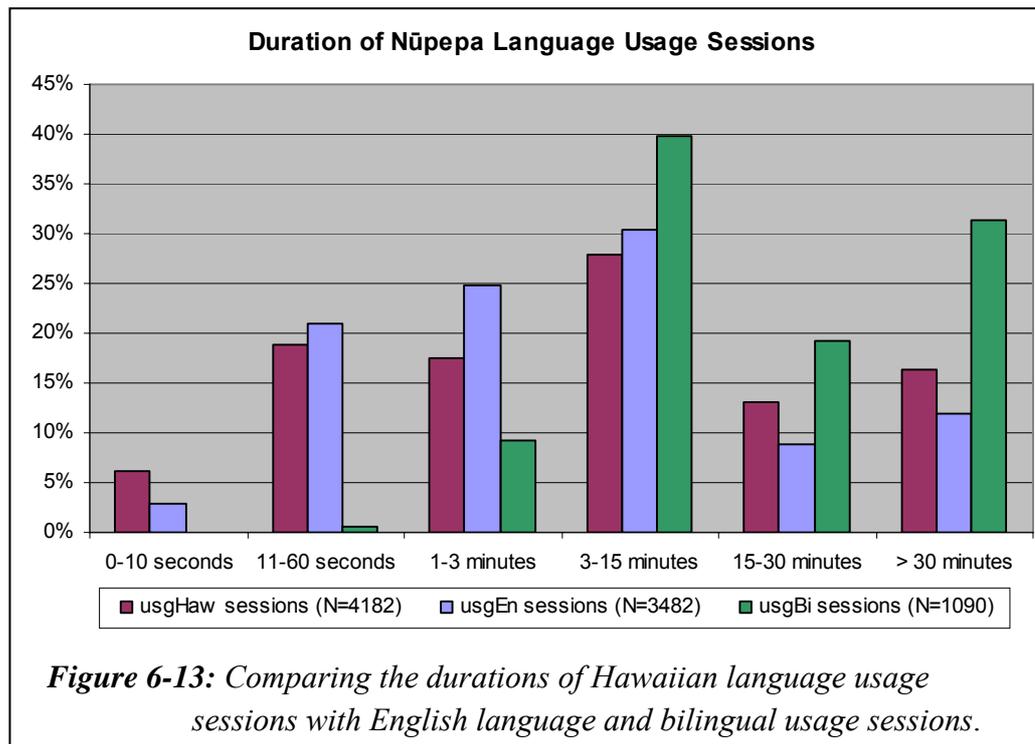


Figure 6-13 shows that usgHaw and usgEn sessions appeared to have similar session durations, while the usgBi sessions were longer. However, when chi square analysis was undertaken on the session lengths of usgHaw and usgEn sessions it generated a p value of $5.0 * 10^{-27}$. This indicates that the difference is statistically significant. UsgHaw sessions had a higher proportion of longer sessions (in the 15–30 minutes and >30 minute category) while usgEn sessions had a higher proportion of shorter sessions (in the session categories 15 minutes or less).

The average usgHaw session duration was 16.3 minutes, the average usgEn session was 12.4 minutes. The usgHaw sessions were, on average, a third longer

than the usgEn sessions but only half as long as the usgBi sessions. By definition a usgBi session must involve at least ten requests so it is expected that these sessions will be long.

How Long in Number of Requests?

Also analysed was the length of usage sessions in terms of number of requests by clients who used the Nūpepa digital library in Hawaiian. The numbers of requests per usgHaw session were grouped into the following divisions; 2 requests, 3–4 requests, 5–8 requests, 9–20 requests, 21–40 requests and 41+ requests. The numbers of sessions per request count category were calculated as percentages and are displayed in *Figure 6-14*.

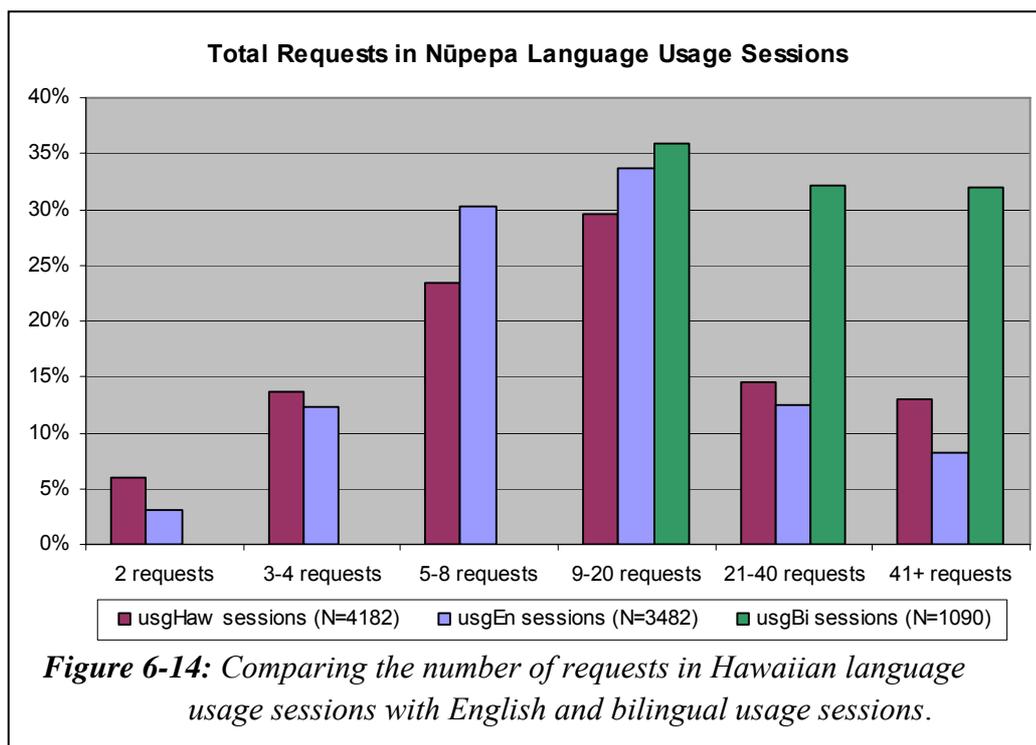


Figure 6-14 shows that there were more short sessions (2 requests and 3–4 requests) and more long sessions (21–40 requests and 41+ requests) in Hawaiian than there were in English. There were proportionally more mid-range (5–8 requests and 9–20 requests) in English than in Hawaiian. UsgHaw sessions had an

average of 21.7 requests and a median of 10 requests per session. The usgEn sessions had an average of 16.4 requests and a median of 9 requests per session.

The statistics for usgBi sessions indicated that these sessions had, on average, over twice as many requests as usgHaw sessions. Again, this high number may simply have occurred because of the criteria for determining a usgBi session; these sessions must be at least 10 requests long.

6.4.2 How Long to Read Interface Pages?

Apparent interface page read times are used to determine how long usgHaw clients took to read pages of the Nūpepa interface. The apparent read times in usgHaw, usgEn, usgBi and in total usage sessions were calculated and descriptive statistics are shown in *Table 6-5*.

Apparent Page Read Times of Interface Pages in Usage Sessions				
	usgHaw	usgEn	usgBi	Total
Total Apparent Reads	4948	5302	3298	13548
Average (seconds)	37.9	37.0	44.9	39.3
Standard Deviation	183.8	179.6	188.1	183.3
Mode (seconds)	3	4	4	3
Shortest session (seconds)	0	0	0	0
Lower Quartile (seconds)	3	5	4	4
Median (seconds)	6	9	9	8
Upper Quartile (seconds)	15	20	21	19
Longest session (seconds)	3492	3489	3412	3492

Table 6-5: Summary statistics of apparent read times of interface pages in usgHaw, usgEn and usgBi sessions.

UsgHaw clients took, on average, 37.9 seconds to read interface pages of the Nūpepa digital library. This average time was over 1/3 quicker than the time it

took expHaw clients to read interface pages in exploratory sessions (61.4 seconds).

The average apparent read time undertaken by usgHaw clients appears similar to the average apparent read time undertaken by usgEn clients. The high standard deviations listed in *Table 6-5* indicate that there is a wide variation in the usgHaw and usgEn apparent read times. The usgBi sessions took the longest time between page requests.

6.4.3 How Many Clients Returned?

Statistics generated from TLA can indicate how often clients returned to the Nūpepa digital library. The statistics are generated under the following assumptions: clients always had cookies enabled, the cookies were not routinely deleted, clients always logged on using the same account on the same computer, and because browser caching was disabled by the Ulukau web server, there was minimal to nil network and local browser caching. While these assumptions were not valid for some users in some sessions, the figures can still give comparative indications on the number of return clients.

The number of clients who undertook return sessions and the number of requests they submitted to the Nūpepa web site were collated. From a total of 6,839 identifiable clients there were 555 (8.1%) clients who undertook return sessions. These clients made 60,042 (30.3%) requests for pages and documents from the Nūpepa web site. While only a small number of clients were identified as return clients, they were responsible for a considerable percentage of the requests.

The return analysis was also undertaken for each client language type. There were 378 (12.9%) usgHaw clients who undertook return sessions, requesting 98,545 (63.6%) pages and documents. There were 270 (8.7%) usgEn clients who undertook return sessions, requesting 15,353 (23.7%) pages and documents. There were also 137 (13.0%) usgBi clients who undertook return sessions requesting 8,617 (16.7%) pages and documents in usgBi sessions.

These figures indicate some important points. First, usgHaw and usgBi clients were the clients most likely to return to the Nūpepa web site, returning one and a half times more often than usgEn users. Second, when usgHaw clients returned to the Nūpepa web site they undertook significantly more requests than usgEn and usgBi clients. There were 21.8 requests submitted per non-returning usgHaw client and 329.3 requests submitted per returning usgHaw client. By comparison, there were 17.5 requests submitted per non-returning usgEn client and 78.2 requests per returning usgEn client. Also there were 46.2 requests per non-returning usgBi client and 153.6 requests per returning usgBi client.

Some clients undertook sessions in different language categories. There were:

- 24 clients recorded in all three session types, usgHaw, usgEn and usgBi
- 79 clients recorded in both usgHaw and usgEn session types
- 42 clients recorded in both usgHaw and usgBi session types
- 61 clients recorded in both usgEn and usgBi session types.

While it was noted that some clients undertook sessions in different language categories, it was not considered significant. For example, the 24 clients who undertook sessions categorised as usgHaw, usgEn and usgBi sessions represented only 0.4% of the total number of unique clients.

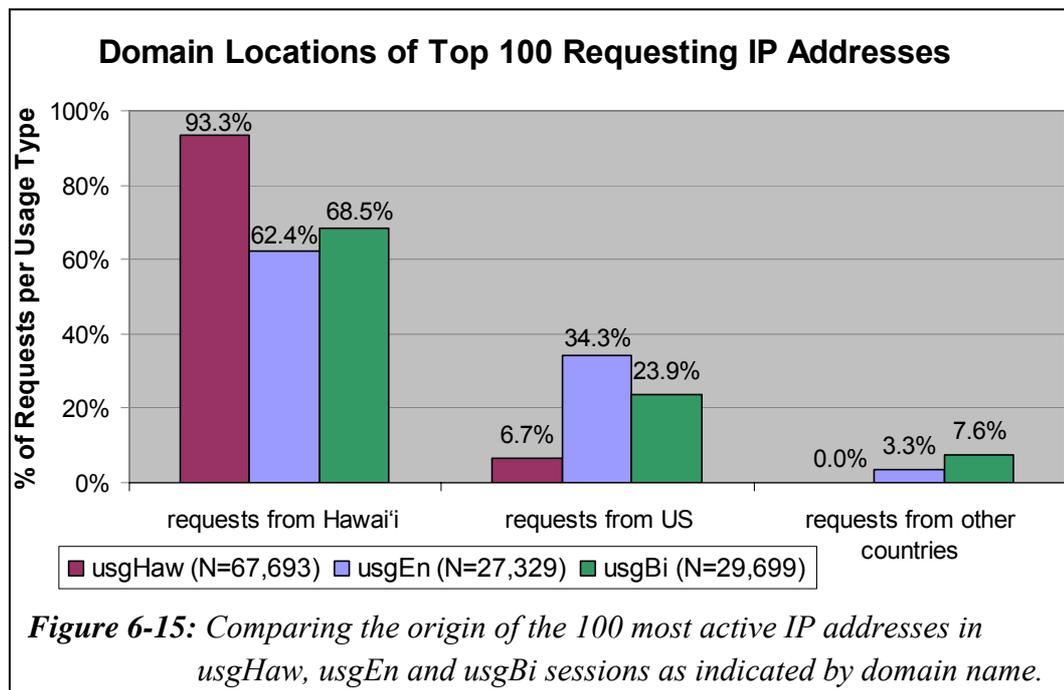
6.4.4 Where Did Usage Sessions Originate?

The TLA process can give information about the origin of requests, as described in §3.4.4. A process called reverse DNS lookup can be used to return information about the computer/network device including domain name and location.

Reverse DNS lookup was undertaken on the 100 IP Addresses which made the most requests in each usage sessions type. The 100 most active usgHaw IP Addresses submitted 67,693 requests, the 100 most active usgEn IP Addresses submitted 17,329 requests, and the 100 most active usgBi IP Addresses submitted

29,699 requests. The 100 usgHaw IP Addresses submitted more than twice as many requests as the usgEn or usgBi IP Addresses.

Results from Reverse DNS lookup indicated three general origins; those IP Addresses that appeared to originate from Hawai‘i, those from the United States and those from other parts of the world. The percentages of requests by location for each usage session type are displayed in *Figure 6-15*.



Most of the requests to the Hawaiian Nūpepa digital library originated from Hawai‘i. Hawaiian usage sessions, at 93.3%, had the highest rate of requests originating from Hawai‘i, suggesting that the resource was being used extensively by local people. Given that over 90% of Hawaiian speakers are of Hawaiian ancestry (Wilson, 2006), these figures suggest solid use of this web site by the indigenous Hawaiian people.

No overseas country made it into the top 100 IP requests in Hawaiian usage sessions. For the English usage sessions there were two IP addresses requesting from Aotearoa (a total of 775 requests) and one from the Netherlands (133 requests). The bilingual usage sessions showed the most involvement from

overseas users; there were two IP addresses requesting from Canada (a total of 295 requests), two IP addresses from Aotearoa (a total of 287 requests), one from Germany (897 requests), one from Holland (430 requests) and one from Japan (360 requests). These figures suggest that users from overseas countries are willing to switch the interface language of the Nūpepa web site between the Hawaiian and English languages.

The question posed in this sub-section was, where did requests by Hawaiian language clients originate? The answer is, they mostly originated in Hawai‘i.

6.4.5 How Often Was the Interface Language Switched?

As the default language of the Nūpepa interface was set to Hawaiian it could be expected that only a small number of the usgHaw sessions would involve a language switch. Only 340 (8.1%) of the total 4,182 Hawaiian usage sessions switched the language of the interface. UsgHaw sessions represented 45.8% of the total usage sessions yet were only responsible for 4.8% of the sessions that undertook interface language switching.

Perhaps somewhat surprising was that only 53.4% of the usgEn sessions switched the interface language. This figure suggests that 46.6% of the usgEn sessions had the Nūpepa web site bookmarked in English.

Slightly more of the language interface switches were a switch from the Hawaiian language to English. The Hawaiian-English interface language switch was evident in 56.5% of the usgHaw session switches, 64.6 % of the usgEn session switches and 52.2% of the usgBi session switches.

6.4.6 When Was the Interface Language Switched?

To determine where the interface language switches were occurring in usgHaw sessions, the sessions were divided into 10 equal portions and the number of switches in each portion was counted. *Figure 6-16* shows the percentages of Hawaiian to English switches for usgHaw sessions and *Figure 6-17* shows the percentages of English to Hawaiian switches for usgHaw sessions. *Figures 6-16* and *6-17* also show the percentages of when switching occurred in usgEn and usgBi sessions.

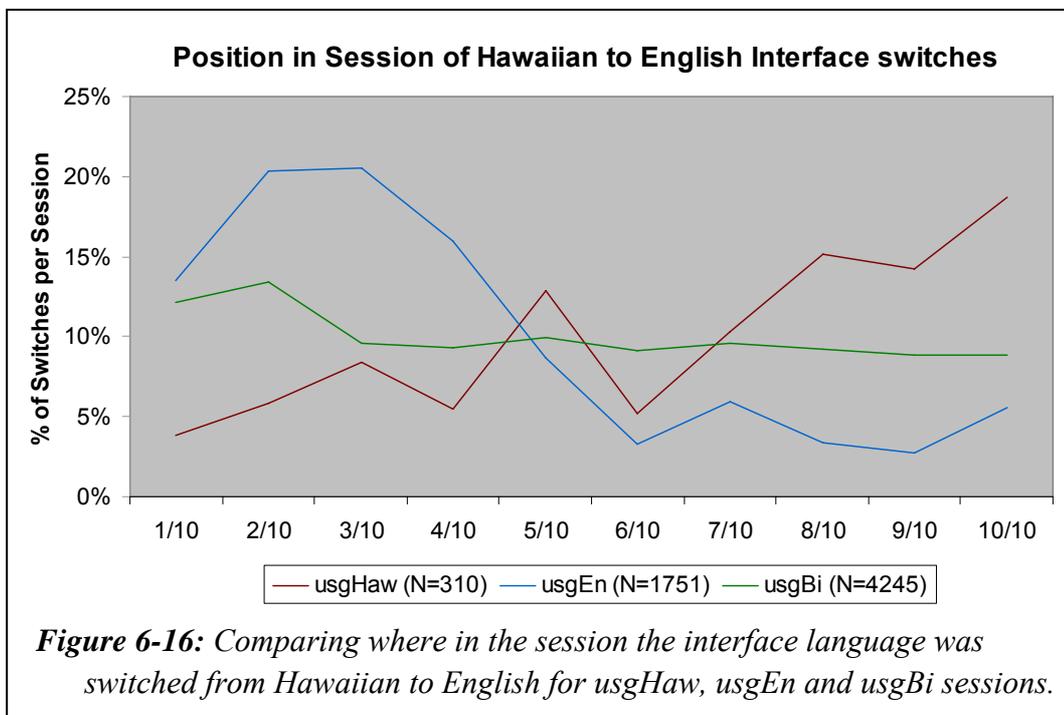
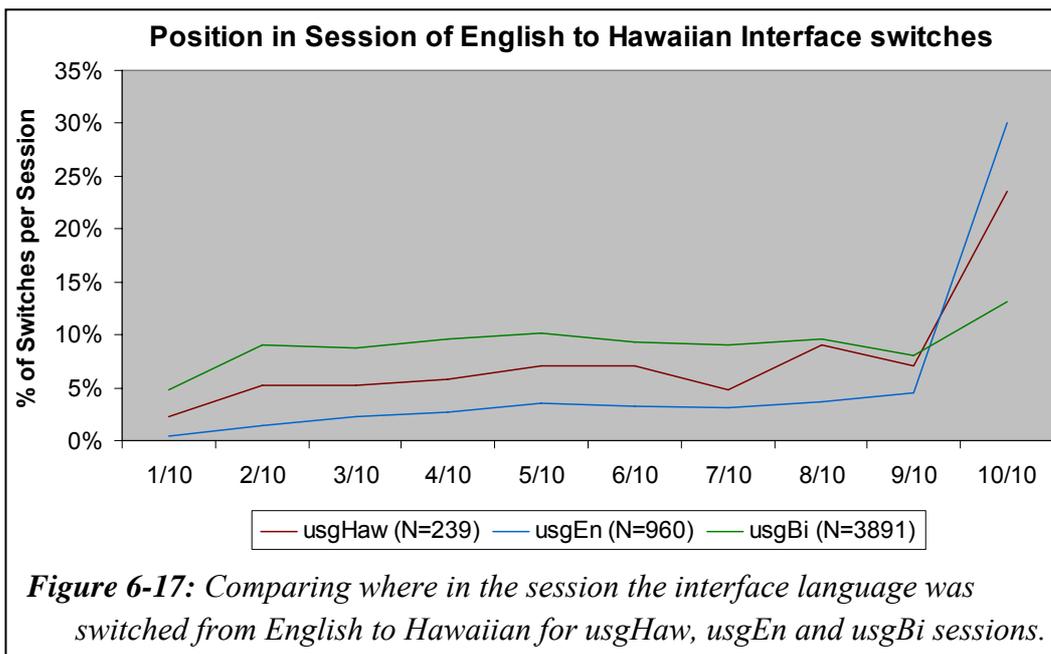


Figure 6-16 shows that when usgHaw clients did switch the language of the interface the switch was most likely to occur towards the end of the session. In contrast, usgEn clients were most likely to switch the interface language from Hawaiian to English at the start of a session, while usgBi clients switched the language of the interface equally throughout the session.

Figure 6-17 shows the interface language switch from English to Hawaiian in usgHaw sessions mostly occurred towards the end of the session. However, it did not occur very often. There were only 239 English to Hawaiian switches in usgHaw sessions, compared with 960 in usgEn sessions and 3,891 in usgBi sessions.



The interface language switch from English to Hawaiian in usgEn sessions also mostly occurred towards the end of the session. However, the interface switch in usgBi sessions occurred at a reasonably constant rate across the whole usgBi session.

6.4.7 Summary—Interface Behaviours

This section analysed how the Nūpepa interface was used by Hawaiian language clients in usage sessions. Six aspects of using the Nūpepa interface in Hawaiian were examined by analysing statistics drawn from TLA.

The statistics revealed that, on average, usgHaw sessions lasted for 16 minutes and consisted of 22 requests. The average apparent read time of an interface page by usgHaw clients was 40 seconds. Almost 13% of the usgHaw sessions were identified as return sessions and these return sessions were responsible for almost 64% of the total usgHaw requests. Most (93%) of the requests submitted by usgHaw clients originated from Hawai‘i. The language of the interface was not switched often by usgHaw clients; when switched, the switch usually occurred towards the end of the session.

This brief summary concludes the examination of how the Nūpepa interface was used in the Hawaiian language. The next stage is to examine how documents were accessed. This analysis begins in the following section.

6.5 Document Access Behaviours

This section analyses how Nūpepa documents were accessed by Hawaiian language clients in usage sessions. By definition (see §2.4.2) if a client requested a document in a session, then that session was regarded as a usage session. Documents available from the Nūpepa digital library consisted of historic Hawaiian language newspaper pages that were served to the client either as a preview image, a mid-size image or a full size image. A small percentage of documents (6.4%) were also available in an extracted text format (see *Figure 6-3*).

TLA was used to provide statistics for seven behaviours exhibited by Hawaiian language clients who accessed documents from the Nūpepa digital library:

- the methods used to access documents in usgHaw sessions
- the document formats favoured in usgHaw sessions
- the Nūpepa periodicals most requested by usgHaw clients
- the publication years most requested by usgHaw clients
- the propensity for documents with associated text files

- the use of the Detach option by usgHaw clients
- the apparent read times of documents by usgHaw clients.

These seven document access behaviours comprise the next seven sub-sections.

6.5.1 How Were Documents Accessed?

As mentioned in §6.1.3 there were three methods by which Nūpepa clients could access documents; they could be requested following a query activity, a browse by series, or a browse by date. When a client requested a document by one of these methods, say a query, and then proceeded to browse documents in the collection, the subsequently browsed documents were recorded as being generated from the query activity. The methods used by usgHaw clients to access documents are displayed in *Table 6-6*, along with the methods used by usgEn and usgBi clients.

	usgHaw		usgEn		usgBi	
Total	48,455	100.0%	20,450	100.0%	24,205	100.0%
From query	13,018	26.9%	12,114	59.2%	12,591	52.0%
From series	31,030	64.0%	4,542	22.2%	5,682	23.5%
From date	4,407	9.1%	3,794	18.6%	5,931	24.5%

Table 6-6: Counts and percentages comparing how documents were accessed by usgHaw, usgEn and usgBi clients.

Clients who used the Hawaiian language (usgHaw clients) preferred to use the series list to access documents. This could indicate that these clients had prior knowledge about specific documents and used the series list to access them directly. Hawaiian language clients (like te reo Māori clients) showed the least interest in accessing documents using a list by date. The query facility might have been used more by usgHaw clients if a higher number of Nūpepa documents had had extracted text files.

UsgEn and usgBi clients preferred to use the query facility to access Nūpepa documents even though only 6.4% of the documents were searchable.

6.5.2 Which Document Formats Were Favoured?

The Nūpepa documents stored in the Nūpepa digital library were made available to the clients in one of four formats; a preview image, a mid-sized image, a full size image in a pdf format and, for a small percentage of documents, an extracted text format. *Table 6-7* lists the total numbers and percentages of these four formats requested by clients.

Formats of Requested Documents						
	usgHaw		usgEn		usgBi	
Total Documents	48,455	100.0%	20,450	100.0%	24,205	100.0%
Preview Image	16,101	33.7%	4,419	22.5%	9,048	38.2%
Mid-sized image	13,148	27.5%	2,082	10.6%	1,870	7.9%
Full Image	2,025	4.2%	535	2.7%	2,518	10.6%
Extracted Text	16,517	34.6%	12,618	64.2%	10,232	43.2%

Table 6-7: Document formats requested by usgHaw, usgEn and usgBi clients.

Preview images and extracted text documents were the formats most requested by usgHaw clients. Approximately 1/3 of the documents requested by usgHaw clients were preview images, 1/3 were extracted text documents and the remaining 1/3 comprised the mid-size and full size images. Considering that only 6% of the documents in the Hawaiian Nūpepa collection had extracted text formats available it is surprising that this format was requested so much.

The extracted text format was the format prominently requested by usgEn clients. This is, perhaps, a result of usgEn clients' preference for using the query facility to access documents. UsgBi clients often exhibited format preferences that were between the extremes of the usgHaw and usgEn clients; usgBi clients preferred

the extracted text format (but not as much as usgEn clients) and also showed the highest preference for preview images.

6.5.3 Which Nūpepa Periodicals Were Favoured?

Identifying which documents of the Nūpepa collection were requested wasn't a straightforward process. The Greenstone software on which the Nūpepa collection was built, had been configured to allocate a unique identifier to each document and subsequently use this identifier when servicing requests for the document. Each unique identifier was created by hashing^{iv} the metadata of the document with any extracted text which had been generated for the document. The unique identifiers were created each time the Nūpepa collection was rebuilt, which happened when new extracted text files or new image files were added to the collection. When the Nūpepa server recorded requests for documents in the Nūpepa web log it recorded the unique identifier of each document. To determine exactly which documents had been requested, the hashing process had to be reversed.

The Nūpepa web log for 2005 was analysed and the unique identifiers extracted. The hashing of unique identifiers for documents was reversed to determine exactly which documents the identifiers were referring to. Unfortunately, changes had occurred in some document details after they were recorded in the 2005 logs. These changes meant that new identifiers were created and the old identifiers, as recorded in the 2005 logs, no longer mapped directly to a document. Of the 93,110 documents requested by clients in 2005, 21,543 (23.1%) did not map directly to a document and therefore could not be used in the document analysis. The remaining 71,567 documents were further analysed to determine exactly which periodicals were accessed by clients who used the Nūpepa digital library in Hawaiian.

^{iv} Hashing is a process where a string of characters has a mathematical process performed on it to produce a unique identifier.

Although 46 periodical titles are currently available on the Nūpepa web site, in 2005 there were fewer than this. No specific records were kept but the web site developers suggest that perhaps there were less than ten periodicals available in 2005. The list of periodicals requested by Nūpepa usage clients, as recorded by almost 77% of the 2005 Nūpepa web log, is displayed in *Table 6-8*.

The most popular periodical was *Ka Nupepa Kuokoa*. With 51,328 requests, this periodical was requested in 71.7% of identified requests, which is perhaps not surprising as the periodical had over 90% of all available pages. The second most requested periodical was *Ka Lama o Hawai‘i* which was requested 8,361 times and accounts for 11.7% of the identified requests. The third and fourth most requested periodicals were *Ka Hoko o Ka Pakipika* at 4,104 requests (5.7%) and *Ka Nonanona* at 3,223 requests (4.5%). No other single periodical was responsible for more than 2% of the identified requests.

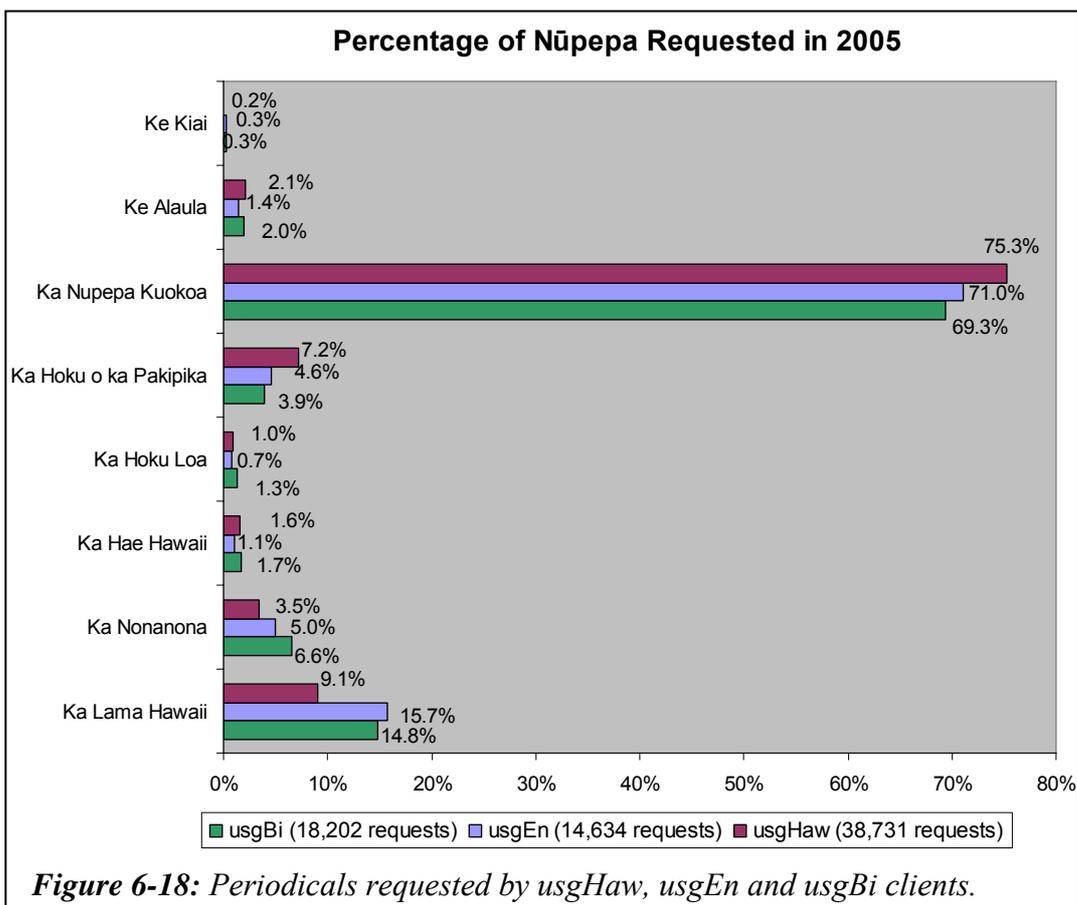
Nūpepa Periodicals Requested in 2005		
Catalogue Number and Name	Requests	Pages Available ^v
001 Ka Lama Hawaii	8,361	109
004 Ka Nonanona	3,223	533
008 Ka Hae Hawaii	1,089	36
009 Ka Hoku Loa	713	104
020 Ka Hoku o ka Pakipika	4,104	325
021 Ka Nupepa Kuokoa	51,328	18,153
024 Ke Alaula	1,370	337
080 Ke Kiai	166	16

Table 6-8: Number of requests per Nūpepa periodical.

^v These were the number of pages available in April 2007. While no accurate records were kept it is suspected that a similar number were available in 2005.

The eight periodicals that were requested in 2005 were then further analysed to see which ones were favoured by usgHaw clients. The percentages of requests per periodical were graphed for usgHaw, usgEn and usgBi clients. These percentages are displayed in *Figure 6-18*.

The periodical *Ka Nupepa Kuokoa* was requested in 75% of requests by clients who used the Nūpepa digital library in Hawaiian. This periodical was slightly (4.3%) more popular with usgHaw clients than it was with usgEn and usgBi clients. The second most favoured periodical by usgHaw clients was *Ka Lama Hawaii* followed by *Ka Hoku o ka Pakipika*.



UsgEn and usgBi clients appeared to have similar preferences for accessing certain periodicals. *Ka Nupepa Kuokoa* was favoured by both client types, followed by *Ka Lama Hawaii*. However, a chi square analysis on the number of

usgEn and usgBi requests for the different periodicals returns a p value of $5.9 * 10^{-21}$. This p value indicates that periodicals requested by usgEn clients were different from those requested by usgBi clients and that the difference was statistically significant.

6.5.4 Which Publication Years Were Favoured?

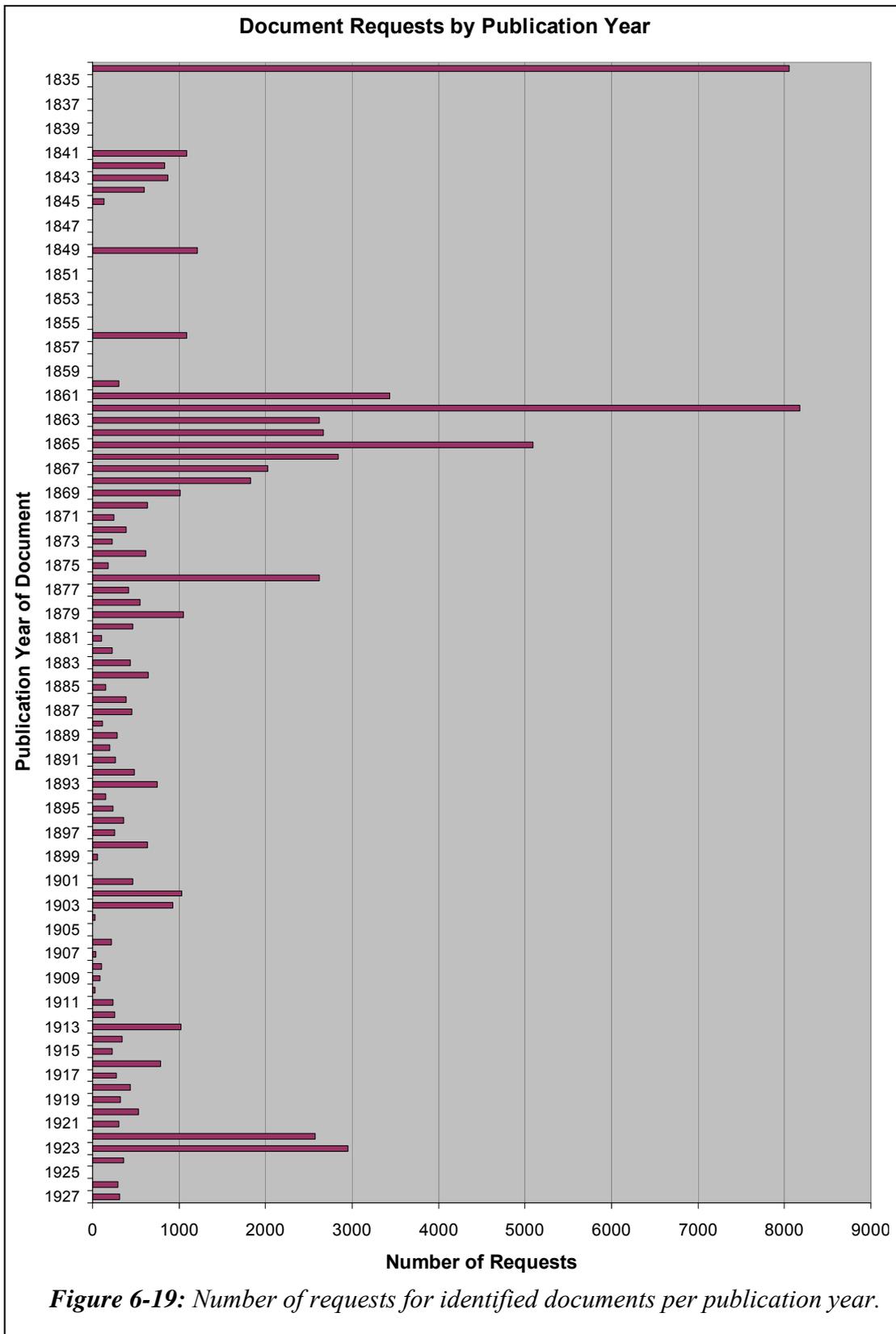
As mentioned in the previous sub-section, 71,567 requested Nūpepa documents could be identified. These identified documents represent 76.9% of the total documents requested by Nūpepa clients. These identified documents were further analysed to determine what year of publication was favoured by Nūpepa clients.

The periodicals in the Nūpepa collection were published over a 93 year time period, from 1834 to 1927. The publication dates of the identified requested documents were collated and are displayed in *Figure 6-19*.

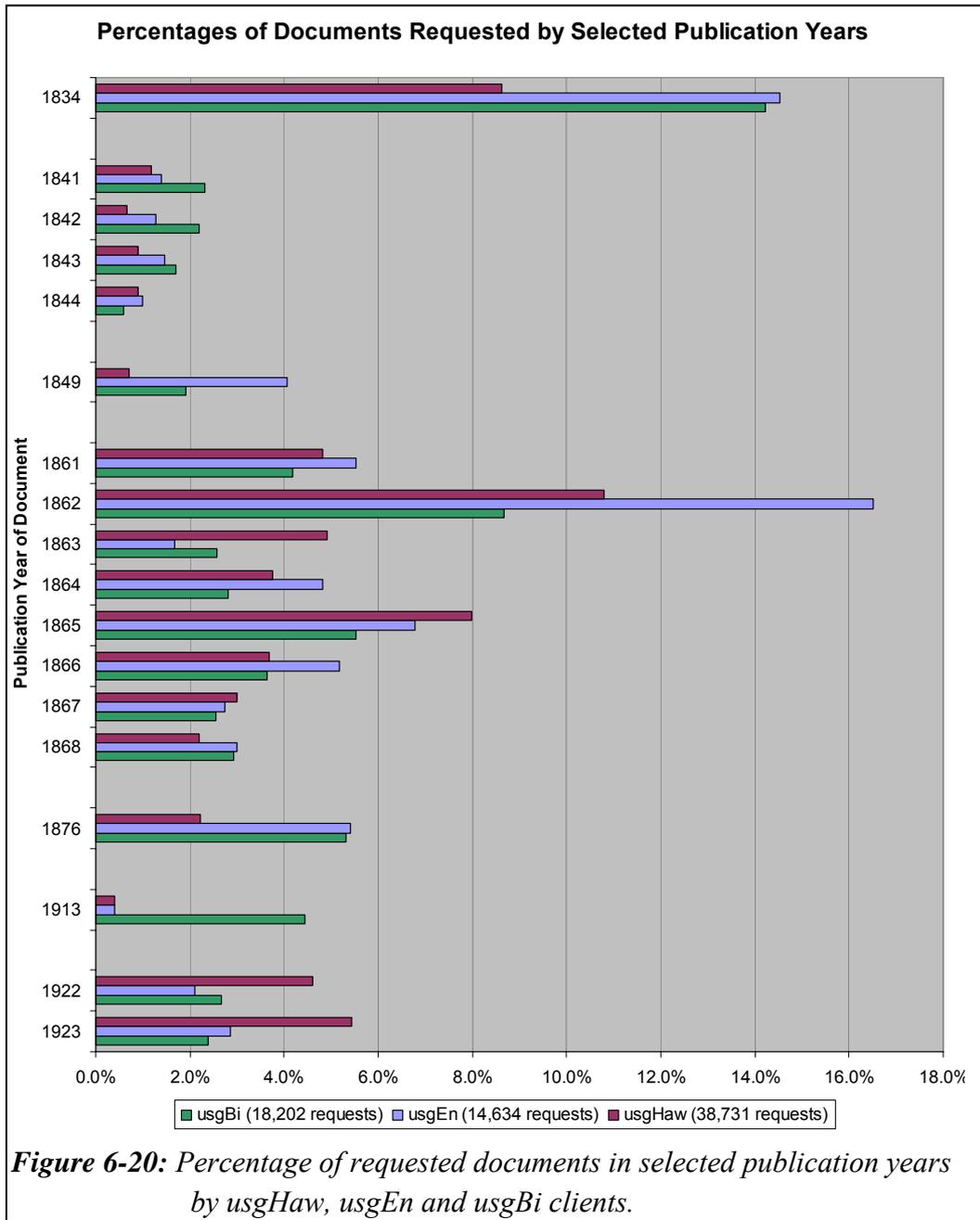
Figure 6-19 shows that the publication dates of the identified documents were most popular in the following years; 1834, 1841–1844, 1849, 1861–1868, 1876, 1913 and 1922–1923. These dates have been listed again in *Figure 6-20* and have been separated into years requested by usgHaw, usgEn and usgBi clients.

Figure 6-20 shows that 1862, 1834 and 1865 were the publication years most favoured by usgHaw clients. Nūpepa documents published in 1862 and 1834 were favoured even more by usgEn and usgBi clients. The documents published in 1865 were most favoured by the usgHaw clients.

Figure 6-20 does show that some publication years were favoured more by usgHaw clients than by other clients. For example 1922, 1923, 1863 and 1865. Other publication years, like 1849 and 1862, were more favoured by usgEn clients while others still, like 1842–1843 and 1913, were more favoured by usgBi clients. The reasons why certain client types displayed higher preferences for request documents in certain years could be due a number of factors; because of the content that was available in those publication years, the different methods clients



used to access documents or because of historic events that had different relevance to clients who used the Nūpepa digital library. Further research is needed to determine exactly why certain client types prefer Nūpepa documents written in certain years.



6.5.5 Documents With Associated Extracted Text Files

Nūpepa documents with associated extracted text files were likely to be requested more often than documents without extracted text files. Having an extracted text file allowed a document to be discovered when a query was submitted.

Of the pages that were able to be identified, 43,964 (61.4%) were Nūpepa documents that had an associated extracted text file (and could be subsequently searched on) and 27,603 (38.6%) were Nūpepa documents that did not have an associated text file. Counts of documents with associated extracted text files requested by usgHaw clients are displayed in *Table 6-9* along with counts of documents requested by usgEn and usgBi clients.

	usgHaw		usgEn		usgBi	
Total	38,731	100%	14,634	100%	18,202	100%
With Text	19,617	50.6%	12,040	82.3%	12,307	67.6%
Without Text	19,114	49.4%	2,594	17.7%	5,895	32.4%

Table 6-9: *The number of requests for documents with and without associated extracted text files by clients, usgEn and usgBi clients.*

Half of the documents requested by usgHaw clients had associated extracted text files. As the documents with associated extracted text files only represented 6.4% of the total Nūpepa documents, it can be stated that clients who used the Hawaiian language (usgHaw) favoured documents with associated extracted texts.

UsgEn clients had a much higher preference (82.3%) for requesting documents with associated extracted text files. This high preference shown by usgEn clients was a result of usgEn clients' dependency on the query facility to locate documents. This dependency meant that 82.3% of the documents requested by usgEn clients were sourced from only 6.4% of the total Nūpepa content.

6.5.6 Was the Detach Option Used?

Nūpepa clients had the option to detach a document. Selecting the detach option opened the document in a new browser window. Having two browser windows available meant that the client could retain a particular document view in one window while continuing to browse/query the Nūpepa digital library in the other window. The detach option was available for the extracted text, the preview image and the mid-size image views but was not available when viewing a full size image.

Data from the Nūpepa filtered log showed 1,989 documents were requested using the detach option in 2005. This number represents only 2.1% of the total documents requested. The detach feature was rarely used by usgHaw clients, only 693 documents were detached by usgHaw clients, a figure that represents only 1.4% of their total document requests.

The detach facility was used most by usgEn clients who requested 792 documents to be detached which still only represented 4.0% of the total usgEn documents requested. The usgBi clients requested 537 detachments at 2.3% of the total usgBi documents requested.

6.5.7 How Long to Read Documents?

Apparent document read times are used to indicate how long clients took to read actual documents (as opposed to interface pages) of the Nūpepa digital library. The average apparent document read times by usgHaw, usgEn and usgBi clients were generated from TLA and are displayed in *Table 6-10*.

(all times in seconds)	usgHaw	usgEn	usgBi	Total	Count
Full Size Image	169.0	133.3	134.2	147.6	4,329
Extracted Text	85.7	87.3	65.0	80.6	37,179
Mid-size Image	73.1	92.9	69.3	75.1	16,439
Preview Image	13.1	19.4	14.3	14.4	29,082

Table 6-10: Average apparent read times of document types by usgHaw, usgEn and usgBi clients.

Table 6-10 shows that, on average, usgHaw clients spent the longest time reading full size images, at almost 3 minutes. This was twice the average time taken to read extracted text documents (85.7 seconds) and mid-size images (73.1 seconds), and thirteen times the average time taken to view the preview images (13.1 seconds). UsgEn and usgBi clients also took the longest times reading full size images and the least times reading preview images.

6.5.8 Summary—Document Access Behaviours

This section analysed how Hawaiian language clients in usage sessions accessed documents from the Nūpepa digital library. Statistics drawn from TLA examined seven document retrieval behaviours exhibited by usgHaw clients.

The statistics revealed that selecting a document from a series list was method favoured by usgHaw clients. This preference contrasted with behaviours displayed by usgEn and usgBi clients of the Nūpepa digital library, and all clients of the Māori Niupepa, who preferred the query facility. The preference for usgHaw clients to use the series list to access documents indicates their awareness that only a small percentage (6.4%) of the documents had extracted text files, and also suggests they had prior knowledge of what documents were available in the Hawaiian Nūpepa digital library.

Statistics from TLA also revealed that preview images and extracted texts were the formats most preferred by usgHaw clients, although mid-size images were also significantly requested. The periodical *Ka Nupepa Kuokoa* was requested in three of four requests by usgHaw. The publication years 1862, 1834 and 1865 were most favoured by usgHaw clients. Half of the documents requested by usgHaw clients had associated extracted text files, an indication that the OCR process is being undertaken on documents relevant to Hawaiian language clients. The detach feature was rarely used by usgHaw clients. Apparent read times indicated that reading full size images took the longest time by usgHaw clients; approximately twice the time of reading mid-size images and extracted text files and thirteen times the time it took to view preview images.

This brief summary concludes the examination into how usgHaw clients accessed Nūpepa documents. The next stage, beginning in the next section, is to examine how queries were submitted by usgHaw clients.

6.6 Query Behaviours

This section analyses the query behaviour of usgHaw clients using the Nūpepa digital library. A query is the action taken by a client who has entered a term or terms into the input box of the Nūpepa search engine and then clicked the ‘Huli/Begin Search’ button.

The query statistics discussed in this section have been generated by the TLA process. Nine query behaviours are analysed in this section:

- the number of queries submitted in usgHaw sessions
- the number of query terms used in usgHaw queries
- the topics queried by usgHaw clients
- the language of query terms
- the altering of query parameters in usgHaw queries
- the use of query operators in usgHaw queries
- the number of documents requested following a usgHaw query
- the formats favoured following a usgHaw query
- the apparent document read times following a usgHaw query.

These nine query behaviours comprise the next nine sub-sections.

6.6.1 How Many Queries Were Submitted?

The number of queries submitted per usgHaw session was grouped into three categories; no queries per session, 1–2 queries per session and 3 or more queries per session. Percentages of queries in each category are displayed in *Table 6-11*.

Almost half of all usgHaw sessions did not involve a query, almost a quarter involved the use of one or two queries per session, while the remaining quarter involved the submission of three or more queries per session.

	usgHaw	usgEn	usgBi	all
Total Sessions	4182	3482	1090	8754
Sessions with no queries	49.0%	35.0%	29.6%	41.0%
1–2 queries (per session)	23.4%	29.4%	13.5%	24.6%
3+ queries (per session)	27.5%	35.7%	56.9%	34.4%

Table 6-11: Queries per session in usgHaw, usgEn and usgBi sessions.

The usgBi clients showed the highest preference for undertaking multiple queries per session, while the usgEn clients showed more use of queries per session than usgHaw clients. The numbers of queries per session submitted by the different client types were plotted and are displayed in *Figure 6-21*.

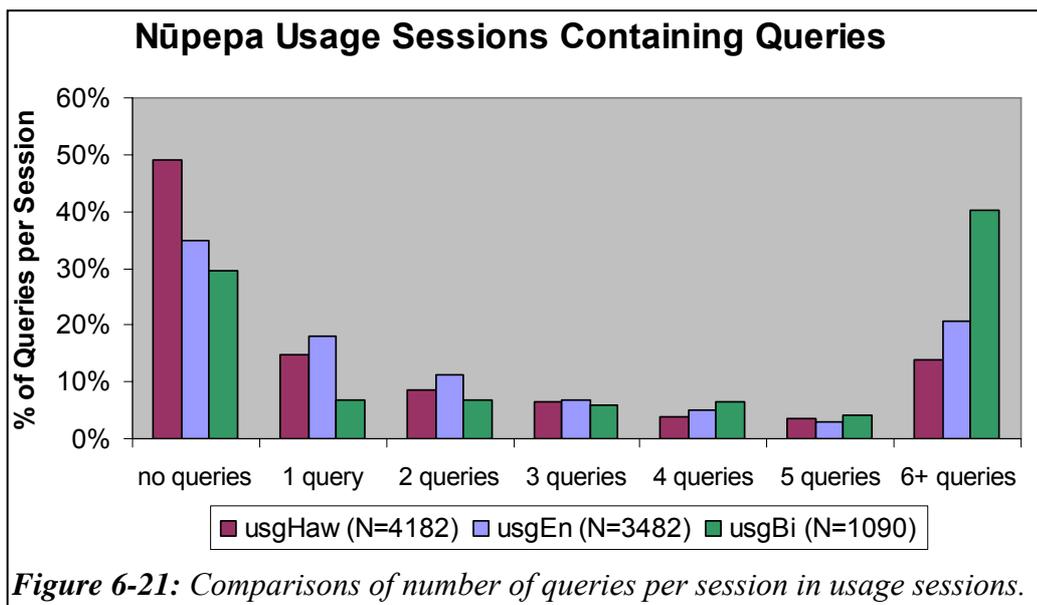


Figure 6-21 indicates again that usgHaw clients undertook the fewest number of queries per session. This figure also highlights that usgEn clients undertook the most single and double queries per session and usgBi clients undertook the most multiple queries per session.

6.6.2 How Many Query Terms Were Submitted?

The number of query terms submitted with each usgHaw query string was counted. Almost two thirds of all usgHaw queries involved just one query term. Approximately one quarter of all usgHaw queries involved two query terms. The remainder involved three or more query terms. The numbers and percentages of query terms per usgHaw query are displayed in *Table 6-12*, along with the numbers of query terms per usgEn and usgBi query.

	usgHaw		usgEn		usgBi	
Total	11,232	100.0%	14,966	100.0%	9,428	100.0%
1 term/query	7022	62.5%	10081	67.4%	5955	63.2%
2 terms/query	2580	23.0%	3553	23.7%	2484	26.3%
3 + terms/query	1630	14.5%	1332	8.9%	989	10.5%

Table 6-12: The number of query terms submitted per query by usgHaw, usgEn and usgBi clients.

In answer to the question posed, most of the time (62.5%), usgHaw queries involved the use of just one query term.

The percentages appear very similar for the three usage session types. However, chi square analysis returns a p value of $6.5 * 10^{-51}$, indicating that there is a significant statistical difference. The usgHaw clients displayed the highest preference for submitting queries with three or more terms (14.5%). The usgBi clients displayed the highest preference for submitting queries with two terms, and the usgEn clients displayed the highest preference for queries with just one term (67.4%).

6.6.3 What Was The Language Of The Query Terms?

The query terms submitted to the Hawaiian Nūpepa search engine were collated. There were 18,478 query terms used in 11,232 queries undertaken by usgHaw clients, 22,026 query terms used in 14,966 queries undertaken by usgEn clients and 14,342 query terms used in 9,428 queries undertaken by usgBi clients.

The most common terms queried are displayed in *Table B-1* which is located in *Appendix B*.

The language of the query terms was noted for the top 100 query terms submitted by usgHaw, usgEn and usgBi clients. Two languages were used, Hawaiian and English. The percentages of Hawaiian and English query terms have been plotted in *Figure 6-22*.

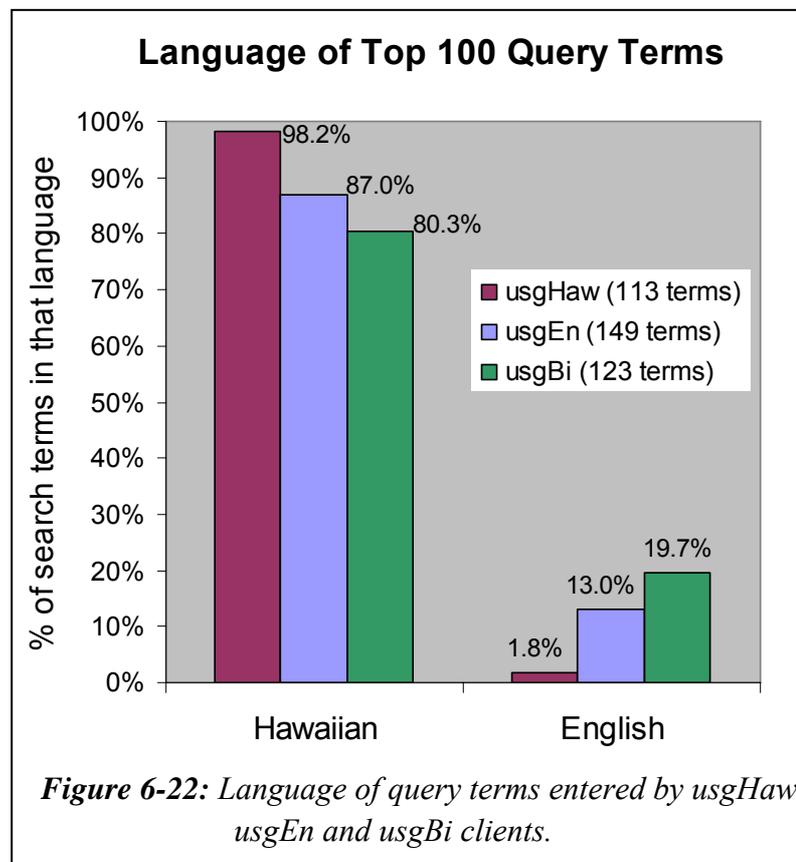


Figure 6-22 shows that Hawaiian was the predominant language of the query terms used in usgHaw queries. There was only one query term in the top 113 usgHaw query terms not in the Hawaiian language; a query for the term ‘*bush*’.

Considering that there are very few English language texts in the Hawaiian Nūpepa digital library it is surprising to see usgEn and usgBi clients using English query terms. The higher percentage of query terms in English by usgBi clients suggests that using English was a deliberate strategy undertaken by usgBi clients to assist in the information retrieval process.

6.6.4 Were Query Parameters Altered?

The Hawaiian Nūpepa digital library offers the user the ability to alter only one query parameter the Some/All parameter. This parameter gives the client the ability to query on *Some* of the query terms or *All* of the query terms. This parameter is available at the centre of the query page as shown in *Figure 6-2*.

The Some/All query parameter was not altered often by usgHaw clients; only 93 (3.2%) unique usgHaw users altered this parameter in 1.4% of the usgHaw queries.

A similar proportion of usgBi clients altered this parameter; 35 (3.3%) unique usgBi users in 0.8% of the usgBi queries. However, even fewer usgEn clients altered the Some/All parameter; 58 (1.9%) unique usgEn clients altered this parameter in 0.7% of the usgEn queries submitted.

6.6.5 Were Query Operators Used?

Counts from the web log indicate that Boolean operators (and, or, not, *, +, -) were not often used in query strings. Less than 0.9% of queries undertaken in usgHaw sessions involved the use of one of these operators, less than 0.3% in usgEn sessions, and less than 0.4% in usgBi sessions.

However, there was significant use of quotation marks to define precise phrase searching. Of the 11,232 usgHaw queries submitted 958 (8.5%) involved the use of quotation marks to specify exact phrase matching. This rate was higher than in usgEn queries (1.6%) and usgBi queries (2.5%). UsgHaw clients entered more query terms when undertaking queries (see §6.6.2); it is probable that those clients would be more likely to use an operator such as quotation marks to assist in defining results in multiple term queries.

6.6.6 Documents Requested Following a Query

Statistics derived from TLA on the Nūpepa log revealed how many documents were requested by usgHaw clients as a direct result of a query. There were 11,232 queries submitted in usgHaw sessions directly leading to the request of 10,702 documents; a return rate of 0.95 documents per usgHaw query. An average of 0.75 documents was requested by usgEn clients following a usgEn query and 1.16 documents requested by usgBi clients following a usgBi query.

The number of Hawaiian Nūpepa documents requested per usgHaw query was almost four times fewer than the number of Māori Niupepa documents requested per usgMi query. This can be attributed to the difference in the Nūpepa/Niupepa results pages. The Hawaiian Nūpepa results page, see *Figure 6-2*, displays two preview lines from the returned document to indicate the context in which the query term is used. In contrast, the Māori Nūpepa results page, *Figure 3-3*, does not show any preview lines but instead just lists the documents in which the query term is located. Clients of the Māori Niupepa collection must open (request) the documents to determine if the returned document is relevant to the query. Opening the documents to determine relevance results in a higher number of Māori Niupepa documents being requested per query as compared to the number of Hawaiian Nūpepa document requests per query.

If the number of Hawaiian Nūpepa documents requested following a query is fewer than the number of Hawaiian queries submitted, it could be an indication that the Nūpepa queries are not returning relevant results to the clients. Given this

scenario, the figures listed above suggest that queries submitted by usgEn clients are returning the least relevant results and queries submitted by usgBi clients are returning the most relevant results. The queries submitted by usgHaw clients, at 0.95 document requests per query, appear to be returning more relevant results than usgEn clients but less relevant results than usgBi clients.

6.6.7 Formats Favoured Following a Query

The document formats requested by usgHaw clients following a query were compared with the document formats requested when a query was not submitted. The percentages are displayed in *Table 6-13*, along with the figures for query and non-query document format requests by usgEn and usgBi clients.

	usgHaw		usgEn		usgBi	
	Non Query	After Query	Non Query	After Query	Non Query	After Query
Total	37,089	10,702	8,481	11,173	12,711	10,957
Extracted Text	21.7%	79.0%	46.0%	78.0%	28.3%	60.5%
Full Size Image	4.2%	4.5%	3.1%	2.5%	9.3%	12.1%
Mid-size Image	34.4%	3.5%	19.4%	3.9%	10.5%	4.9%
Preview Image	39.7%	12.9%	31.6%	15.6%	51.8%	22.5%

Table 6-13: Comparing types of documents requested not involving a query with those requested following a query from usgHaw, usgEn and usgBi clients.

Table 6-13 clearly shows that formats favoured when a query was not submitted were quite different from those favoured following the submission of a query. When a query had not been submitted almost 75% of the requests from usgHaw clients were for mid-size and preview images. However, when a query was submitted over 75% of the requests were for the extracted text documents.

UsgEn and usgBi clients displayed similar behaviours. When a query was not submitted the mid-size and preview images were preferred; when a query was submitted the extracted text documents were preferred.

6.6.8 Document Read Times Following a Query

The apparent read times of documents following a query were compared with the apparent read times of documents when no query was involved. The non-query and query average apparent read times by usgHaw clients are displayed in *Table 6-14*, along with the non-query and query average apparent read times by usgEn and usgBi clients.

(times in seconds)	usgHaw		usgEn		usgBi	
	Non Query	After Query	Non Query	After Query	Non Query	After Query
All Documents	49.7	96.9	56.3	85.6	35.0	73.6
Extracted Text	66.2	104.8	61.5	98.8	37.8	79.8
Full Size Image	167.2	174.8	148.9	120.3	123.4	144.0
Mid-size Image	71.9	115.4	95.8	81.9	60.2	91.7
Preview Image	40.5	23.1	51.6	18.8	25.5	17.1

Table 6-14: Comparing apparent read times of documents not involving a query with read times following a query for usgHaw, usgEn and usgBi clients.

Table 6-14 shows that, overall, the apparent read times of documents by usgHaw clients did change following the submission of a query. Apparent read times by usgHaw clients of extracted text documents, of full size images and of mid-size images were all longer following the submission of a query. However, the apparent read times of preview images by usgHaw clients were shorter following the submission of a query.

The apparent read times of documents by usgEn and usgBi clients also changed significantly following the submission of a query and were similar to those changes displayed by usgHaw clients.

6.6.9 Summary—Query Behaviours

This section analysed how queries were undertaken on the Nūpepa digital library by clients who used the Hawaiian language. Statistics drawn from TLA examined nine query behaviours exhibited by usgHaw clients.

The statistics revealed that almost half (49%) of all usgHaw sessions did not involve a query. When a query was submitted by usgHaw clients, half of them involved the use of one or two queries per session while the other half involved the submission of three or more queries per session. Where sessions contained a query, more than half (62.5%) of the usgHaw queries contained just one query term. While only 14.5% of usgHaw queries contained three or more query terms, this did represent a much higher preference by usgHaw clients for using multiple query terms than was shown by usgEn and usgBi clients. UsgHaw clients predominantly used Hawaiian language query terms (for a listing of the query terms submitted see *Appendix B*). The Some/All query parameter was altered only rarely by usgHaw clients, which was also the case in the use of Boolean operators in usgHaw queries. However, quotation marks were used in 8.5% of the usgHaw queries to define precise phrase matching.

Query statistics also revealed that usgHaw clients requested an average of 0.95 documents for every query submitted. UsgHaw clients who submitted queries primarily requested extracted text documents, while those who did not submit queries primarily requested mid-size and preview images. The apparent read times of documents by usgHaw clients increased for all document types following a query, apart from preview documents. The apparent read times of preview documents by usgHaw clients was much shorter following the submission of a query.

Usage of the Hawaiian language by clients who actively *used* the Nūpepa digital library has been discussed in three parts; usgHaw interface behaviours, usgHaw document access behaviours and the section summarised here, usgHaw query behaviours. Two further Nūpepa usage questions are posed: how did indigenous Hawaiian language usage differ from non-indigenous English language usage of the Nūpepa digital library; and how was the Hawaiian language bilingually used with the English language by clients of the Nūpepa digital library? These two questions will be addressed in the next two sections.

6.7 Hawaiian Language Usage/English Language Usage

The purpose of this chapter has been to define how the Hawaiian language was used by clients of the Nūpepa digital library in 2005. There is also merit in examining how this usage compares with English language usage of the Nūpepa digital library. If differences can be identified then they can be incorporated into digital library design to further assist indigenous language use, and non-indigenous language use, of a digital library.

A comprehensive usgHaw and usgEn comparison can be undertaken by examining the statistics listed in each sub-section of sections 6.4–6.6. The intention of this section, §6.7, is to summarise the key differences and similarities between how the Nūpepa digital library was used in Hawaiian and how it was used in English. This section comprises two parts; behaviour differences in expHaw and expEn exploratory sessions and behaviour differences in usgHaw and usgEn usage sessions.

Behaviour Differences in Exploratory Sessions

Nūpepa exploratory sessions in Hawaiian were slightly, but significantly, longer than exploratory sessions in English. The expHaw sessions had an average longer

duration (1.63 minutes compared with 1.56 minutes) and a higher average number of requests per session (3.0 requests compared with 2.3 requests) than expEn requests. The apparent interface page read times of expHaw clients was statistically similar to those by expEn clients. These figures imply that expHaw clients often requested one more interface page than the expEn clients before deciding not to use Nūpepa digital library.

Behaviour Differences in Usage Sessions

The Hawaiian Nūpepa digital library was used more by usgHaw clients than by usgEn clients. UsgHaw clients were responsible for 45.8% of the total usage requests while usgEn clients were responsible for only 28.8% of the requests (25.4% were submitted by usgBi clients). On average, usage sessions in Hawaiian were longer than those in English. The usgHaw sessions were longer in duration (16 minutes compared with 12 minutes) than the usgEn sessions and involved a larger average number of requests (22 requests compared with 16 requests).

UsgHaw clients took a similar time to read the interface pages (in usage sessions) as usgEn clients. This similarity was also noted in the exploratory sessions. The usgHaw clients had a higher percentage of return sessions (12.9% compared with 8.7%) and submitted more requests per return client (329 requests compared with 78 requests) than usgEn clients. A higher rate of return sessions and return session requests implies that the Nūpepa digital library was more relevant to usgHaw clients than to usgEn clients. Just over 93% of the usgHaw requests originated from within the Hawaiian Islands compared with only 62% of the usgEn requests.

It is difficult to compare usgHaw and usgEn language switching behaviours as the default language of the Nūpepa digital library was set to Hawaiian for all of 2005. Thus, many usgHaw clients did not need to switch the interface language while in theory, all of the usgEn clients did need to. Further, if usgHaw clients did entertain some form of language switching then they were often redefined as usgBi clients.

UsgHaw clients preferred to request documents from a series list, usgEn clients preferred to request documents by undertaking a query. This, in turn meant that usgHaw clients had a higher preference for requesting the mid-size or preview image formats whereas usgEn clients had a higher preference for requesting the extracted text formats. As usgEn clients were requesting a higher proportion of extracted text documents they were also requesting documents that mostly had associated extracted text files. Documents with associated text files were requested in 82.3% of the documents requested by usgEn clients compared with only 50.6% of the documents requested by usgHaw clients.

There appeared to be a small but significant difference in the periodicals and years of publication requested by usgHaw and usgEn clients. UsgHaw clients showed higher preferences than usgEn clients for requesting documents from *Ka Nupepa Kuakoa* and *Ka Hoku o ka Pakipika* and for the years 1863, 1865 and 1922–1923. Alternatively usgEn clients showed higher preferences than usgHaw clients to request documents from the *Ka Lama Hawai‘i* and documents in the years 1832, 1849, 1862 and 1876.

As mentioned above, usgEn clients preferred to use a query to access documents. This occurred even though only 6.4% of the Nūpepa documents were available to the search facility. UsgEn clients undertook more queries per session than usgHaw clients; however, these queries usually involved fewer query terms per query. UsgEn clients used English query terms in a small percentage of queries (13.0%) whereas usgHaw clients undertook almost no queries in English (1.8%). UsgEn clients, like usgHaw clients, made little use of query parameters. While usgHaw clients used quotation marks to specify exact phrase matching in 8.5% of queries, usgEn clients only used this operator in 1.6% of queries. As usgHaw clients showed a higher propensity to submit queries with multiple terms it also follows that usgHaw clients would make more use of this phrase matching operator.

UsgEn queries did not lead to as many document requests as usgHaw queries; there was an average of only 0.75 documents requested following a usgEn query compared with 0.95 documents requested following a usgHaw query. When the

rate of document requests per query is lower than one it suggests that many queries are not returning relevant results.

The submission of a query affected the types of documents that were requested. When a query was not submitted usgHaw clients preferred to request full size images. However, following a query usgHaw clients had a much higher preference for extracted text documents, while the preferences for full size and preview images reduced by almost a half. UsgEn clients did not show the same differences following a query. UsgEn clients had the highest preference for extracted text documents whether a query was submitted or not. The only difference following a query by usgEn clients was a reduction in the preference for accessing preview images which subsequently led to an increase in the preference for extracted text documents.

The submission of a query also affected the apparent read times of documents. In usgHaw sessions the apparent read times following a query decreased for full size images, remained the same for extracted text documents but increased for preview images. In usgEn sessions, the apparent read times increased following a query increased for full size images, decreased for extracted text documents and stayed the same for preview images.

UsgEn clients showed a strong preference for switching the interface language from Hawaiian to English usually at the start of a session. When usgHaw clients switched the interface language from Hawaiian to English they did it towards the end of a session. Both client types were recorded switching the interface language from English to Hawaiian, though not to the same extent, and both client types generally made this switch at the beginning or end of a session.

6.8 Bilingual Usage

The analysis of the Māori Niupepa digital library has suggested that bilingual use of the Niupepa digital library was a strategy undertaken by a small number of clients to increase the effectiveness of information retrieval (see §3.8). Statistics generated from TLA of the Hawaiian Nūpepa digital library indicate that this strategy was undertaken by a large rather than small number of clients. There were 1,090 unique Nūpepa clients categorised as usgBi clients and these clients were responsible for 25.4% of the requests to the Nūpepa collection. There were almost as many requests undertaken by usgBi clients (50,359 requests) as by usgEn clients (57,154 requests).

By definition (see §2.4.3) usgBi sessions must contain at least 10 requests. Perhaps because of this, usgBi sessions had almost twice the duration and twice the number of requests as usgHaw and usgEn sessions. Likewise, the number of documents requested per session was much higher for the usgBi sessions; 22.2 documents per session while usgHaw sessions requested 11.6 documents per session and usgEn sessions requested 5.9 documents per session.

There were similar proportions of return usgBi and usgEn clients at 13% of the total clients. However the return usgBi clients were only responsible for 17% of the total usgBi requests whereas the return usgHaw clients were responsible for 64% of the total usgHaw requests.

Document access behaviours exhibited by usgBi clients tended to fall between those behaviours displayed by usgHaw clients and those behaviours displayed by usgEn clients. The methods used to access documents, the preferred document formats, the favoured periodicals and publication years, the preference for documents with extracted texts, the use of the detach facility and the apparent read times by usgBi clients were all behaviours that fell between usgHaw and usgEn client behaviour.

The manner in which usgBi queries were undertaken was different from other clients. UsgBi clients undertook more queries per session (8.7) than usgHaw (2.7) or usgEn clients (4.3). UsgBi clients also used more English in query terms (19.7%) than usgHaw (1.8%) or even usgEn clients (13.0%). UsgBi clients appeared to have the most effective querying strategies as they recorded the highest rate of document requests (1.2) following a query. UsgBi clients were also the least interested in accessing the extracted text formats following a query. While usgBi clients did request extracted text formats in 60% of the documents requested following a query, this percentage was 20% lower than requests by usgHaw and usgEn clients. UsgBi clients seemed more interested than usgHaw or usgEn clients in the image form of Nūpepa documents following a query.

In summary, it appears that the Hawaiian language was used with the English language in a deliberate strategy to assist in information retrieval and to request a relatively high number of documents. The way the indigenous language was used bilingually in the Hawaiian Nūpepa digital library is remarkably similar to the way the indigenous language was used bilingually in the Māori Niupepa digital library. The striking difference was that the Hawaiian Nūpepa digital library was used considerably more bilingually (25.4% of requests) than the Māori Niupepa digital library (2.3%).

6.9 Summary

Statistics from Transaction Log Analysis (TLA) have been used in this chapter, to determine how the Hawaiian language was used by clients of the Nūpepa digital library in 2005. This chapter began by giving background information about the Hawaiian Nūpepa digital library and then described, with the aid of screenshots, how clients could access documents within the digital library.

Requests to the Nūpepa digital library were then categorised into sessions and language of sessions to enable investigation whether the Hawaiian language was used by clients of the Nūpepa digital library. Results indicated that the Hawaiian Nūpepa digital library was used predominantly in the Hawaiian language. Almost half of the usage requests were defined as Hawaiian language usage sessions (usgHaw), a little over one quarter were defined as English language usage sessions (usgEn) and the remaining quarter were defined as sessions where the clients used the interface in both Hawaiian and English (usgBi).

Analysis was undertaken to determine how Hawaiian language clients explored the Nūpepa interface but then did not go on to use it. Exploratory requests comprised only 1.2% of the total requests to the Nūpepa digital library. Hawaiian language exploratory clients, in a similar manner to English language exploratory clients, only made two or three requests per session and most sessions were completed in less than half a minute.

Analysis of how Hawaiian language clients actually used the Nūpepa digital library was undertaken using statistics generated from TLA. These statistics showed that usage sessions undertaken by usgHaw clients averaged 16 minutes and 22 requests. The Hawaiian language usage analysis was discussed in three categories; usgHaw interface behaviours, usgHaw document access behaviours and usgHaw query behaviours.

UsgHaw interface behaviours showed the average apparent read time of an interface page by usgHaw clients was 38 seconds. At least 13% of the usgHaw

sessions were return sessions and usgHaw clients who returned to use the Nūpepa digital library were responsible for almost 64% of the total usgHaw requests. Fourteen out of fifteen requests submitted by usgHaw clients originated from the Hawaiian Islands. UsgHaw clients did not switch the language of the interface often; less than 1 usgHaw session in 12 involved an interface language switch.

UsgHaw document access behaviours indicated that usgHaw clients preferred to select documents from a series list. This preference is contrary to behaviours displayed by usgEn and usgBi clients of the Nūpepa digital library, and by all clients of the Māori Niupepa digital library, who all showed the highest preference for selecting documents by using the query facility. Statistics also revealed that preview images and extracted texts were the formats most preferred by usgHaw clients, although mid-size images were also significantly requested. Half of the documents requested by usgHaw clients had associated extracted text files. The detach feature was not used significantly by usgHaw clients. The average apparent document read time by usgHaw clients were almost three minutes for a full size image, almost 1.5 minutes for an extracted text file, just over 1.2 minutes for a mid-size image and 40 seconds for a preview image.

UsgHaw query behaviours revealed that almost half (49%) of all usgHaw sessions did not involve a query. When a query was submitted, half involved the use of 1–2 queries per session while the other half involved the submission of three or more queries per session. UsgHaw clients entered more query terms per query than usgEn and usgBi clients. UsgHaw clients predominantly queried terms written in the Hawaiian language. The Some/All query parameter was not used significantly by usgHaw clients; however quotation marks were used in 8.5% of the usgHaw queries to define precise phrase matching.

Query behaviours also revealed that usgHaw clients requested an average of 0.95 documents for every query submitted. UsgHaw clients who submitted queries primarily requested extracted text documents, while usgHaw clients who did not submit queries primarily requested mid-size and preview images. UsgHaw clients took longer to read all types of documents following a query, apart from the preview images.

Activities undertaken by Nūpepa clients who used Hawaiian were compared with activities undertaken by Nūpepa clients who preferred English. ExpHaw clients, on average, requested one more interface page per session than the expEn clients. UsgHaw sessions were a third longer than usgEn sessions in duration and in number of requests. More usgHaw clients than usgEn clients returned to use the Nūpepa digital library and the usgHaw clients were responsible for more requests. Almost all usgHaw requests originated from the Hawaiian Islands while a significant proportion (38%) of usgEn requests originated from outside of the Hawaiian region. UsgHaw clients preferred to select documents from the series list and favoured accessing preview and mid-size images. UsgEn clients preferred to select documents from the results of a query and favoured accessing extracted text documents. These preferences led to usgHaw and usgEn differences in document access behaviours and query behaviours (as has been discussed in §6.8).

The bilingual use in Hawaiian and English was analysed in the previous section. This section revealed that the usgBi clients were a significant Nūpepa client group responsible for a quarter of all requests to the Nūpepa digital library. UsgBi sessions were longer than usgHaw and usgEn sessions, both in duration and in number of requests. UsgBi clients demonstrated document access behaviours that were between those displayed by usgHaw and usgEn clients. However, the usgBi clients were the client group most focused on using the query facility to access documents. UsgBi clients exhibited unique searching behaviours, such as using a high percentage of English search terms, to achieve accurate results from queries.

This chapter has charted Hawaiian language usage by clients of the Nūpepa digital library in 2005 and has clearly shown how the Hawaiian language was used by clients to retrieve information from the Nūpepa digital library. Like te reo Māori, the Hawaiian language was being used by choice without prompting or coercion. This chapter has provided a second example of how an indigenous language was used by clients of a digital library.

The previous four chapters have examined *how* indigenous languages were used by clients of digital libraries. TLA has been used to give comprehensive analysis of what indigenous language usage occurred on two separate bilingual indigenous

digital libraries in 2005. The statistics have been used to give a picture of what happened but not why it happened. Chapter 7 will use two methods to gain insights into *why* indigenous languages were used by clients of digital libraries.

7 Why was the Indigenous Language Used by Digital Library Clients?

Statistics derived from Transaction Log Analysis were used in the previous four chapters to investigate *how* indigenous languages were used by clients of digital libraries. The results showed that use of indigenous languages did occur suggesting that digital libraries provide an electronic medium for the use of an indigenous language. Why did clients choose to use the indigenous language when working with the digital? The purpose of this chapter is to seek feedback from clients of digital libraries to suggest answers to this question.

Two methods were used to obtain client feedback; a web questionnaire and client interviews. The web questionnaire was placed on the Māori Niupepa web site for 250 days in 2006. The questionnaire was presented to clients when they requested the home page of the Niupepa digital library. There were 60 responses to the web questionnaire. Client interviews were undertaken with users or potential users of the Niupepa and Nūpepa digital libraries. Those interviewed comprised 3 clients who used the Māori Niupepa interface in te reo Māori, 8 clients who used the Hawaiian Nūpepa interface in Hawaiian and 19 clients who used the Māori Niupepa interface in English. Although the sample size was small some relevant responses were collected which give important insights into indigenous language usage of digital libraries.

The question of why indigenous languages were used by clients of digital libraries will be answered in two parts; why indigenous language speaking clients used

these particular digital libraries and why they used them in an indigenous language. Feedback indicates the content of the digital library attracted indigenous language clients, along with the fact that this content was in the indigenous language. The indigenous language was used by clients of digital libraries because the indigenous language matched the content of the digital library, was the default language of the digital library and because the clients wanted to use the indigenous language as often as possible.

7.1 Web Questionnaire Methodology

As mentioned in the introduction, two methods were used to obtain client feedback. This section describes the first method, the web questionnaire. Web questionnaires, also known as on-line questionnaires, are a widely used technique to solicit feedback from clients who by their very nature cannot be questioned directly. They are considered easier, more efficient and less error prone than paper based questionnaires (Bradburn, Sudman, & Wansink, 2004). A full description is given of the web questionnaire, followed by an explanation of how it was presented to Niupepa clients in 2006.

7.1.1 Web Questionnaire Description

The questionnaire specifically sought to determine the contexts in which the indigenous language was used. As it was only used on the Māori Niupepa web site, it collected information only about the contexts in which te reo Māori was used by Niupepa clients.

The web questionnaire is shown in *Figure 7-1*.

Web Questionnaire Form

Three Minute Niupepa Feedback

Tēnā koe. Tēnei te inoi atu ki a koe kia tautokohia mātou i a mātou e rangahau kōrero ana e pā ana ki te kohikohinga nei.
Please take a few minutes to answer some anonymous questions about your usage of this website.
For questions about this feedback form please email Te Taka Keegan (tetaka at cs.waikato.ac.nz). Ka nui te mihi.

Ka whakamahia e koe tēnei kohinga mō te aha? (pāwhiritia ngā whakautu e hāngai ana)
What do you use this collection for? (select which answers apply)

Rangahau kōrero e pā ana ki te iwi *Researching tribal information*
 Rangahau kōrero e pā ana ki a koe, ki tō whānau rānei *Researching personal or family information*
 Rangahau ngā āhuatanga o neherā *Researching historical practices*
 Rangahau kupu/āhuatanga o te reo *Research on words/language*
 Rapu kōrero whakamāori/whakapākehā *Search for translation texts*
 Whakangahau/rapu kōrero hei pānui noa *Fun/recreational reading*
 Ētehi atu, whakamōhio mai *Other, please specify*

He aha ngā tumomo mahi i tirohia ai e koe tenei kohikohinga?
For which types of industry are you accessing this collection?

Mahi mō te iwi *Tribal research*
 Rangahau i te whare wānanga *Academic research*
 Mahi ako *Teaching*
 Kawanatanga *Government*
 Ture *Law*
 Paoho *Media (tv, radio, print)*
 Rongoa *Medicine*
 Whaiaro motuhake *Personal research*
 Ētehi atu, whakamōhio mai *Other, please specify*

He aha ngā painga o te whakamahi i te kohinga nei?
What do you find most rewarding about using the collection?

He aha ngā mea kāore i te pai o te whakamahi i te kohinga nei?
What do you find most frustrating about using the collection?

I a koe e whakamahia ana te kohikohinga nei i te nuinga o te wā ka noho te reo o te atanga ki tēwhea reo?
As you are using this collection, the language of the interface is mostly set to which language?

Māori
 English
 Aua *Not sure*

He kōrero i tua atu?
Any other comments?

Done

Figure 7-1: A screen shot of the web questionnaire form

The web questionnaire was a parallel language form with texts written in te reo Māori first, followed by English equivalents in italics. The form began with a greeting in both te reo Māori and English, followed by six questions and a button to submit the form. The six questions asked were:

- *What do you use this collection for?*
- *For which types of industry are you accessing this collection?*
- *What do you find most rewarding about using the collection?*
- *What do you find most frustrating about using the collection?*
- *As you are using the collection, the language of the interface is mostly set to which language?*
- *Any other comments?*

One question not asked was “Why do you use the digital library in te reo Māori?” This was seen as a leading question that might bias the clients’ responses to being overtly positive or overtly negative towards te reo Māori.

The web questionnaire was titled the ‘Three Minute Niupepa Feedback’. It was given this title in the hope that clients would realise that answering the form would take only three minutes of their time and, thus, encourage more clients to complete it.

7.1.2 Web Questionnaire Presentation

The web questionnaire was designed to be as unobtrusive as possible to clients of the Niupepa web site. The questionnaire would only appear to those clients who had enabled cookies on their web browsers and requested the Niupepa home page for a second time. It was a pop-up window that could simply be ignored, and full access to the Niupepa digital library could continue. Once a client had completed and submitted the questionnaire it would not re-appear to that client.

Two cookies were created to handle the client interaction with the questionnaire. A cookie called *niupepasurvey_visited* and *niupepasurvey_blocked* were loaded onto a client’s computer when they first visited the Niupepa web site. Both

cookies were set to 'true'. The *niupepasurvey_visited* cookie was set to expire on the 1 January 2007, while the *niupepasurvey_blocked* cookie was set to expire when the client's current session with the Niupepa web site expired. When the client started a second session on the Niupepa web site the states of these cookies were examined; if the *niupepasurvey_visited* cookie was set to 'true' and the *niupepasurvey_blocked* cookie was set to 'false' then the questionnaire was displayed to the client. Once the client had submitted the survey the *niupepasurvey_blocked* cookie was set to 'true' with an expiry date of 1 January 2007. As both cookie states were now set to 'true' the questionnaire was not redisplayed to that client on subsequent visits to the Niupepa site.

The web questionnaire was activated on the Puka and Thor web servers on 20 of February 2006 and deactivated 3 November 2006. In total, the web questionnaire was active for exactly 250 days and sixty responses were received. Their responses are analysed in detail in §7.2.

7.2 Web Questionnaire Responses

Sixty responses to the web questionnaire were received. Some further responses were recorded when the questionnaire was being set up and tested; however, as these were not submitted by actual clients of the Niupepa digital library they were not included in the questionnaire analysis. There were 48 questionnaires logged by Puka (the national web server) and 12 by Thor (the international web server).

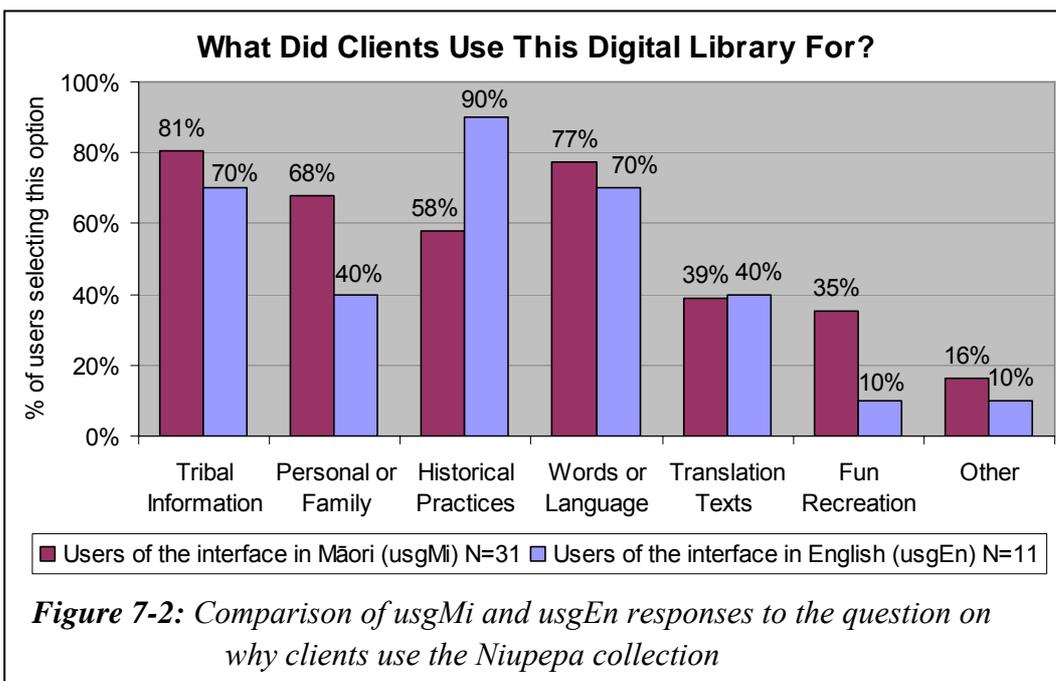
While sixty responses were recorded, not all of the responses included answers for all of the questions. There were ten nil responses, that is, responses that just had the time stamp recorded. A nil response is a result that would occur if the client scrolled through to the end of the form without answering any of the questions and clicked on the *Tukua/Submit* button. Two of the nil responses were from national responses; the other eight were from international responses. The international clients appeared to show a strong resistance to this method of collecting data whereas the national clients appeared prepared to respond.

Of the 60 clients that responded to the questionnaire, 31 stated that they used the interface in te reo Māori (usgMi), 11 that they used the interface in English (usgEn), 6 that they were unsure which language they used the interface in and the remaining 10 gave no indication of which interface language they preferred to use.

As shown in *Figure 7-1* the web questionnaire asked six questions. The responses received for each of these questions are detailed in the next six sub-sections. This section then concludes with a summary of the information deduced from the responses to the web questionnaire.

7.2.1 What Did Clients Use the Digital Library For?

The first question asked on the web questionnaire was, ‘*what do you use this collection for?*’ As shown in *Figure 7-1* clients could choose from six predetermined options or use the seventh option to specify an unlisted option. The percentages of responses selected for each client type have been graphed in *Figure 7-2*.



The Niupepa digital library was used by usgMi clients predominantly for *researching tribal information*, the option selected by 25 of 31 usgMi respondents. This response was closely followed by the *words or language* option which was selected by 24 of 31 usgMi respondents. *Researching personal or family* information also rated highly with 21 of 31 usgMi respondents.

It appears that usgMi clients had a stronger preference for using the Niupepa digital library in a wider range of categories than usgEn clients. Most apparent were the stronger preferences for the categories *fun/recreational reading* (35% compared with 10%) and *researching personal or family information* (68% compared with 40%). The usgEn clients displayed a higher preference for researching *historical practises* (90% compared with 58%).

Seven respondents selected the *other* option four of which entered what the other reasons were. These were entered as:

"Work Purposes"

"He mea tino ataahua, he mea koi ki te hinengaro"

[something very beautiful, something to sharpen the mind]ⁱ

"whakamohio au ko wai au i roto i tenei ao hurihuri"

[to know who I am in this changing world]

"Hei hanga rauemi"

[to create resources]

Figure 7-2 suggests that English language clients preferred to use the digital library to find information on historical practices, while te reo Māori clients preferred to locate tribal information, personal information and information about (reo Māori) words and language. However, there was a low response rate number; there were only 31 responses classified as usgMi and only 11 responses classified as usgEn. This count is too low for any firm conclusions to be drawn.

ⁱ All translations in this chapter are by the author.

7.2.2 Which Research Sectors Used the Digital Library?

The second question was about the research sectors (or industry types) that accessed this collection. As displayed in *Figure 7-1*, eight options were provided plus the option to select ‘*other*.’ The responses collected have been calculated as percentages and are displayed in *Figure 7-3*.

The usgMi respondents on average selected two responses to this question, while usgEn respondents generally selected three. This lower amount of responses selected suggests that the usgMi respondents were using the digital library, and possibly reo Māori resources in general, across fewer sectors than usgEn respondents.

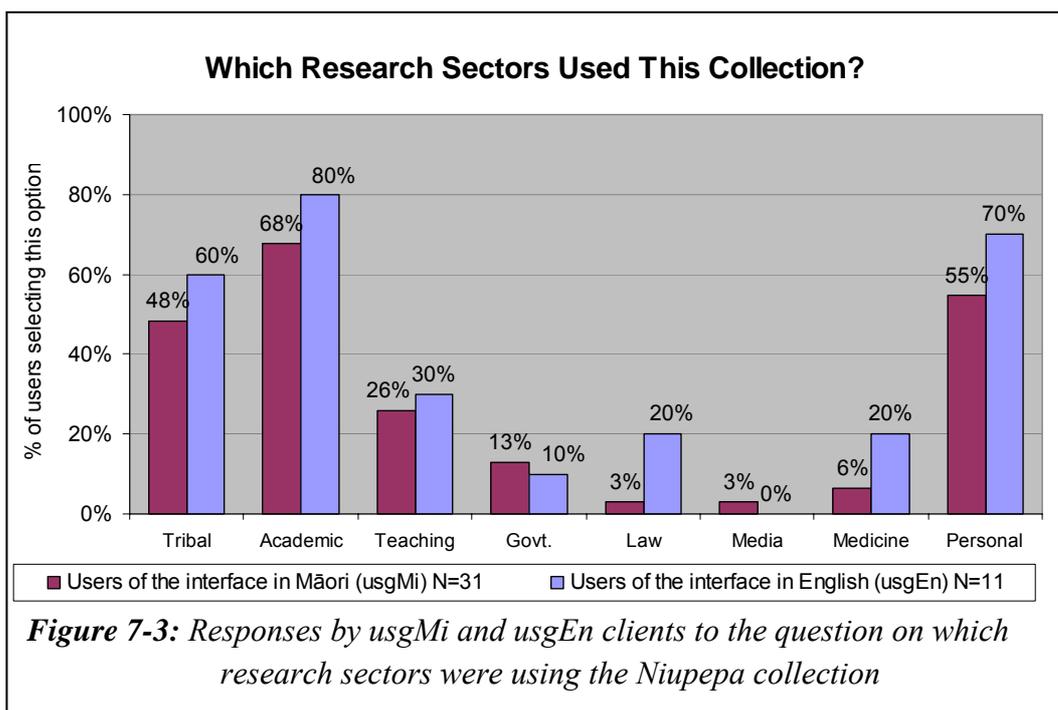


Figure 7-3 indicates that three research sectors were prominent users of the Niupepa digital library in te reo Māori. These fields were *academic* (21 of 31 responses), *personal* (17 of 31 responses) and *tribal* (15 of 31 responses). While it appears that the Niupepa digital library was not used often in te reo Māori by clients in other sectors, it was difficult to draw solid conclusions because of the low number of responses received to the questionnaire.

Four respondents selected the *other* option. These were entered as:

“Hei rauemi ki te whakapai i te Wikipedia”

[As a resource to enhance the Wikipedia]

“Mahi mo te whanau me te hapu”

[work for the family and sub-tribe]

“Treaty claims research”

“Assignment”

7.2.3 What Did Clients Find Most Rewarding?

There were 44 responses for the first open ended question, ‘*what did you find most rewarding about using the collection?*’ Of these, 31 respondents classified themselves as usgMi clients, 11 respondents classified themselves as usgEn clients and a further two respondents gave no indication.

Both usgMi clients (24 of 31) and usgEn clients (7 of 11) mentioned that the newspaper content of the Niupepa digital library was the most rewarding aspect of the digital library. A comment on the content was by far the most common response to this question. Some of the descriptions describing the value of the content were poetically written:

“Ko te mea nui, kua hikoi au i te wa e ora ai a tatou
kuia, a tatou koroua”

[The important thing is that I have walked in the times that my ancestors
were alive]

“Ko te hohonutanga o nga korero kai runga i nga parirau
o nga manu nei.”

[The deepness of the information delivered on the wings of these birds (the
newspapers were often referred to as messenger birds)]

“Te wairua o a tatou kuia, a tatou koroua i te wa e ora ai ratou, e tipu ake ratou e tangata whenua ana ratou nga mea o te ao maori hei taonga tuku iho mo tatou. E kore au e ngaro, he kakano i ruia mai i Rangaiatea.”

[The spirit of our ancestors in the times they were alive, when they grew up, when they were masters of the Māori world passed on as a treasured heirloom for us. I can never be lost, I am a seed sown from Rangiātea.]

UsgMi responses to the web questionnaire indicated that the second most rewarding aspect of the Niupepa digital library was te reo Māori. Te reo Māori was selected in 11 of 31 usgMi responses. Comments were made regarding two areas; te reo Māori written in the newspapers and te reo Māori used in the digital library interface. Examples of comments made in the web questionnaire include:

“Ko te maamaa me te ataahua o te reo o neheraa”

[The clarity and beauty of the old language]

“Firstly, reading the thoughts of my tipuna who wrote letters to the newspapers, through which I have gained an understanding of the choices that they made. Secondly improving not only my understanding of the Maori language but more importantly learning my own dialect as written by my own people!!”

The ability to undertake queries was the third aspect that clients found rewarding, as mentioned by six usgMi respondents and four usgEn respondents. Also mentioned was the availability of the information, having additional information (English abstracts and newspaper commentaries) and the browsing features offered on the web site. Two usgMi respondents also used this question to thank the builders of the digital library for making the Niupepa collection available.

7.2.4 What Did Clients Find Most Frustrating?

There were 42 responses recorded to the second open-ended question which asked, ‘*what did you find most frustrating about using the collection?*’ There were 28 respondents who classified themselves as usgMi clients, six as usgEn clients and a further eight respondents who gave no indication.

The most common response to this question was nothing, that is, there was nothing frustrating. This response was given by 10 (of 28) usgMi clients and 3 (of 6) usgEn respondents. Typical responses were:

“Horekau he raruraru”

[No problems]

“Ka pai te katoa, ki ahau.”

[It’s all good to me]

“Kare kau, he taonga, he waka hua [huia], he mea miharo, he mkea[mea] ataahua...”

[Nothing, a treasure, a repository of treasures, something amazing, something beautiful]

“...kotahi anake te raru, kaore e whai atu te haora, me te ra, me te wiki, me te marama, me te tau ki te noho ana ki te panui pukapuka (only one problem, not enough hours in the day, the week, the months or the year to sit down and feel the wairua o enei korero)”

[...to sit down and feel the spirit of these texts]

There were 18 responses from usgMi clients that did list difficulties. The most common, mentioned by 7 respondents was the fact that the OCR texts contained errors. There were also issues with navigation (2), searching (2), typing errors on the home page (2), the large number of results following a search (2), printing (1), the layout (1), the highlighting (1) and missing newspapers (1).

7.2.5 What Language Was Used by Clients Who Responded To the Questionnaire?

Analysis of this question is in two parts. First, which interface language the respondents stated they used, and second, which language the respondents actually used to answer the open-ended questions in the web questionnaire.

There were 48 responses to the question about which interface language clients used. There were 31 respondents (64.6%) who stated that they used the Niupepa digital library with the interface language set to reo Māori (usgMi), 11 (22.9%) who stated that the interface language was mostly set to English (usgEn) and 6 (12.5%) who stated that they were unsure which language the interface was set to.

For the open-ended questions 32 mostly used te reo Māori when answering the questionnaire. Of these, 27 stated that they used the reo Māori interface, one used the English interface and four seemed unaware of which interface they used as they chose the '*Aua, Not sure*' option. The client who responded in te reo Māori but stated they used the English interface was the only example of a potential reo Māori client choosing to use the interface in English. The respondent actually stated that this was the first time that he/she had used the collection. This is evidence to suggest that perhaps bilingual clients will use an interface in English until they become familiar with it and then use the interface in the indigenous language.

Three clients stated that they used the Niupepa interface in te reo Māori but responded to the web questionnaire in English which indicate that these clients were not highly literate in te reo Māori. This, in turn, could suggest that the Māori interface was still suitable for use by clients with little reo Māori knowledge.

It should also be remembered that while the web questionnaire was written in two languages, te reo Māori was the prominent language seen. It appeared on the first line and began each question. Te reo Māori was not written in italics. Perhaps if the more commonly spoken English language had been the prominent language

more responses may have been written in English, perhaps fewer nil responses and, possibly, more responses received over all.

7.2.6 Further Comments From the Web Questionnaire

The final question solicited further comments. This question was answered by 32 respondents; 23 responded in te reo Māori and 9 in English. The overwhelming response was positive; thanks and praise for the Niupepa digital library and encouragement for those who had made it available on the Web. There were 27 out of the 32 respondents who gave thanks, praise or encouragement. Some examples of these responses are listed below;

“E kore e mutu aku mihi ki a kotou {koutou} ki te rauemi nei me o taua nei tipuna nana nei te kupu i tuku ki te komata o te rangi hei tahu mo taku whare korero.”

[My unending thanks to you people and to this resource and to our ancestors who sent their words to the exalted heavens as a backbone for my house of learning]

“He tino taonga ngaa niupepa nei. Moo ngaa mea peenei i a au nei, horekau atu he huarahi e whiwhi ai au ki eenei koorero. Ka mihi kau atu.”

[These newspapers are a valuable treasure. For people like me, we have no other avenue to access this information. I extend my thanks.]

“Very valuable. Keep up the good work!”

Responses to the final question also included some suggestions and some advice. The suggestions and advice were either written separately or included with some responses expressing gratitude. These responses included a suggestion to use a PDF format for presenting the Niupepa documents, sourcing other newspapers to add to the collection, sourcing other material to add to the collection, correcting the syntax errors and a call not to translate the niupepa documents into English.

One final aspect noted was that usgMi clients appeared less likely than usgEn clients to state a problem with the digital library and more likely to praise. When a problem was stated by usgMi clients, often the English language was used.

7.2.7 Summary—Web Questionnaire Responses

Using the web questionnaire to elicit user feedback was an important exercise because it generated a wide range of responses from clients who used the Niupepa digital library in te reo Māori. The feedback raised some important issues about why clients have used the Niupepa digital library in te reo Māori, issues that otherwise may not have come to light.

It appears that te reo Māori was used for a wide range of reasons but across a relatively small range of industries. The primary reasons for accessing the digital library in te reo Māori were for undertaking tribal research, research on te reo Māori and/or research on personal or family information. A significant number of usgMi respondents also selected *researching historical practices*, *search for translation texts* and *fun/recreational reading* as reasons for accessing the digital library. The range of industries that appeared to be accessing the Niupepa digital library was limited to academic institutions, tribal organisations, with personal researchers also using the resource.

In contrast, the Niupepa digital library appeared to be used in English for fewer reasons by a wider range of industries. UsgEn clients were most likely to use the digital library to research historical practices and while interest was shown in other topics it was generally lower than that shown by usgMi clients. A wide range of industries appeared to be accessing the Niupepa digital library in English; academic institutions, tribal organisations, along with personal researchers were selected by more than half of the usgEn respondents but a significant number also selected the education (teaching) and law and health (medicine) industries.

Both usgMi and usgEn clients stated that the content in the historic newspapers was the most rewarding part of the Niupepa digital library. UsgMi users showed a

strong interest in the written reo Māori used in the newspapers themselves. UsgMi clients also indicated that they were pleased that the interface of the Niupepa digital library was in te reo Māori. It is not surprising that usgEn users did not show the same interest in a language that, perhaps, most of them did not readily comprehend.

OCR errors in the reo Māori extracted texts were listed as the most frustrating issue by the usgMi clients. Te reo Māori appeared to be an important issue for the usgMi clients. Navigation, searching and printing were mentioned more by usgEn than by usgMi clients.

In general usgMi clients seemed less likely than usgEn clients to offer negative comments about the digital library unless specifically asked, as in the fifth question. Where negative comments were made they were often made in English. UsgMi clients also appeared to offer more praise and note their appreciation for the Niupepa digital library, often using poetic phrases to do so.

7.3 Client Interview Methodology

As mentioned in the introduction to this chapter, two methods were used to obtain client feedback. This section describes the second method, client interviews. A full description is given of how the client interviews were conducted, followed by a description of how clients were selected to be interviewed.

7.3.1 Client Interview Procedure

Interviews were conducted with clients, or potential clients, of the Māori Niupepa digital library and the Hawaiian Nūpepa digital library. In total 16 interviews were conducted, involving between one and three participants per interview. Three interviews were with usgMi clients, five with usgHaw clients and eight with usgEn clients. A total of 30 people were interviewed.

The format, comprising three distinct stages, was consistent for all interviews. In the first stage the purpose of the research was explained; those being interviewed were then shown, and had explained to them, a Bill of Rights and a Research Consent Form, which they were asked to sign if they were happy to be interviewed. The Research Consent Form was also signed by the interviewer/researcher. A copy of the Bill of Rights and the Research Consent Form are listed in *Appendix C Interview Forms*.

The second stage of the interview involved an observational study. The participants were observed using the Māori Niupepa digital library or the Hawaiian Nūpepa digital library. Experienced clients were asked to demonstrate the ways in which they used the digital library. If participants had not used the Niupepa/Nūpepa digital library before, the purpose of the digital library was explained and then they were given some predefined tasks to complete. These tasks would require them to seek and retrieve information from the digital library. The interviewer generally remained in the background to observe and record, influencing the outcome as little as possible. At times the interviewer was required to explain the set tasks or to clarify why certain actions were taken. The observational stage of the interview lasted approximately 20 minutes.

The third and final stage of the interview involved asking the participants some prepared questions selected from the following:

- Why do you use the Niupepa collection?
- What types of information do you mostly seek?
- By which methods do you prefer to find information?
- By which methods do you prefer to browse information?
- Which language do you use when entering query terms?
- Do you alter the default language of the web site?
- Does the default language of the web site alter your language use?
- How often do you switch languages in a session?
- Do you have any issues, comments or thoughts about this digital library?
- Do you have any issues, comments or thoughts about indigenous language use of a bilingual web site?

Answering these questions took approximately 30 minutes. Participants were then given a small gift for taking part in the research.

Each interview stage was reviewed and the actions undertaken carefully documented. Items recorded included: the time spent on each particular activity; the success or otherwise of the activity; how the interface was used to access the information; what preferences were altered; what query operators and parameters were used; what difficulties were encountered, and how the indigenous language was used in the activities. Any comments or responses made during the observation and question/answer phases were carefully transcribed.

7.3.2 Finding Clients To Be Interviewed

In theory, identifying Māori Niupepa clients or Hawaiian Nūpepa clients for client interviews should have been difficult as no usage information was kept that distinguished precisely who the clients were. In practice, however, finding these clients was straightforward.

Many te reo Māori clients had provided feedback, both verbal and via email, to the Niupepa web site developers. These clients were contacted and agreed to participate in the client interviews. In fact, more te reo Māori clients were willing to be interviewed than time allowed.

Dr. Robert Stauffer, of Alu Like Inc.ⁱⁱ in Hawai‘i, was contacted and asked if he was aware of any Hawaiian Nūpepa digital library clients who would make suitable participants for this study. In a short time five interviews had been arranged with eight people who were known to Dr. Stauffer, all of whom used the Hawaiian Nūpepa digital library in the Hawaiian language.

Sourcing participants who used (or could use) the digital library in English was also not difficult. At the time participants were being sought, a lecturer of a

ii Alu Like Inc. was one of the organisations responsible for creating the Hawaiian Nūpepa digital library. A more detailed description of them is given in §6.1.2.

computer science undergraduate course circulated a request asking if his students could be interviewed in a usability study. A component of the students' course required them to experience the role of an interviewee. Consequently eight interviews were conducted with 19 computer science students, all of whom were asked to use the Māori Niupepa digital library. Although none of the students had used the digital library previously they were still able to offer some insightful comments. All had previous experience with using electronic resources and most had used electronic resources in a language other than English. Many of the students were bilingual or multilingual and most were visiting students from overseas, which meant their opinions arose from a wide range of perspectives.

Table 7-1 gives a summary of participants in the 16 interviews. The first three interviews were with usgMi clients who used the Māori Niupepa digital library in te reo Māori. The next five interviews were with usgHaw clients who used the Hawaiian Nūpepa digital library in Hawaiian. The final eight interviews were with computer science students who were introduced to the Māori Niupepa digital library and used it mostly in English. A description and summary of each interview is presented in the following section.

Summary of Users Interviewed				
	Digital Library	Usage Language	Participant	Nationalities
usgMi #1	Niuepepa	reo Māori	1 male	Aotearoa Māori
usgMi #2	Niuepepa	reo Māori	1 female	Aotearoa Māori
usgMi #3	Niuepepa	reo Māori	1 male	Aotearoa Māori
usgHaw #1	Nūpepa	Hawaiian	1 male	USA
usgHaw #2	Nūpepa	Hawaiian	1 female	USA
usgHaw #3	Nūpepa	Hawaiian	1 female 1 male	Hawai'i ⁱⁱⁱ
usgHaw #4	Nūpepa	Hawaiian	2 females	Hawai'i
usgHaw #5	Nūpepa	Hawaiian	1 female 1 male	England, Hawai'i
usgEn #1	Niuepepa	English	2 males	Aotearoa, South Africa
usgEn #2	Niuepepa	English	1 female 1 male	China
usgEn #3	Niuepepa	English	2 males	Aotearoa, Taiwan
usgEn #4	Niuepepa	English	2 males	China, Zimbabwe
usgEn #5	Niuepepa	English	2 males	Austria, Germany
usgEn #6	Niuepepa	English	3 males	Germany
usgEn #7	Niuepepa	English	1 female 2 males	Aotearoa
usgEn #8	Niuepepa	English	1 female 2 males	Austria, Aotearoa, Aotearoa Māori

Table 7-1: Summary information about the clients who were interviewed using the Niuepepa or Nūpepa digital libraries

iii While native Hawaiians could be classified as of USA nationality they are listed in this table as Hawai'i to distinguish indigenous Hawaiian people from non-indigenous residents of Hawaiian.

7.4 Client Interview Responses

This section gives a summary of each interview conducted with clients of the Māori Niupepa digital library and the Hawaiian Nūpepa digital library. The summary was taken from the interview documentation created when the video of each interview was analysed.

The setting of each interview and how the interview took place is explained. Some background information is given about each participant so that their responses can be considered in context. For example, a young participant who has a lot of experience using online electronic resources will give different responses from an older participant with little experience using online resources. Multilingual and bilingual participants are likely to respond differently to monolingual participants. Anonymity was maintained at all times during the interview process and in subsequent analysis; none of the interviewees are able to be identified from this summary.

This section is divided into four sub-sections. The interviews with usgMi clients are described first, followed by the interviews with the usgHaw clients and then with the usgEn computer science students. The final sub-section summarises the information detailed in the usgMi, usgHaw and usgEn interviews.

7.4.1 Interviews With UsgMi Clients

Three usgMi participants were individually interviewed as clients of the Niupepa digital library in te reo Māori. All were Māori; one was a native speaker of te reo Māori, the other two were second language learners and all spoke English fluently. Two were men and one was a woman and all had used the Niupepa digital library to various degrees. The interviews took place in Hamilton, Wellington and Christchurch (all cities in Aotearoa) at their places of work. Two interviews took place in October 2005 and the other in May 2006. The

participants remain anonymous. All interview recordings on all forms of media are scheduled to be destroyed by December 2007. Brief summaries of the three interviews follow.

UsgMi Interview #1

This participant was an academic in the field of te reo Māori, Māori customs and Māori history, in the over 50 years age bracket and with little experience using online collections. The participant had, however, used the Niupepa collection, perhaps six times a year for four years. The interview took 50 minutes in total, of which 29 minutes were usage observation and 21 minutes were spent responding to the prepared questions.

The participant was observed undertaking a topic query and examining six results in detail, with both the extracted text files and full size images being studied. The participant noted results on a piece of paper to be followed up at a later date and then submitted another query by entering the title of a periodical into the search box and pressing Rapu (Search). The results were examined, along with commentary information for the particular newspaper. Some time was spent trying to resize the font to make reading the texts easier. Some more documents were read until the conclusion of the observation session.

The prepared questions elicited the following responses: the Niupepa digital library was used to find suitable texts for translation exercises and to undertake tribal research; full text search was the method used to find most information because "it's easy"; reo Māori query terms were used as they matched the topic of the information being sought; initially the digital library was used in both English and te reo Māori but as "time went by I did not need to use that [switch to English] button"; the digital library was used in whatever language the interface default was set to.

UsgMi Interview #2

A tribal researcher working for a Māori tribal authority, in the 31–40 age bracket the participant was classified as a medium user of the Niupepa digital library, using it perhaps once a week for three years. Other online collections had also been used to the same extent. The observation part of the interview lasted for 18.5 minutes and responses took a further 10 minutes.

The participant gave query activity examples and entered a recent query for a particular name. The relevance of the results was discussed. Another query was undertaken, this time on a dialectal term, showing relevant results to the tribe for whom the participant was working. A third query was undertaken, for a historical letter, and then a fourth query on the participant's name; in both cases the results returned were relevant. The participant showed how the extracted text files were copied and used as material for the tribal web site and as sources for tribal educational material and traditional greetings. The participant also used the greetings in personal correspondence.

The participant then responded to the prepared questions: mostly the Niupepa digital library was used to find traditional tribal information that could then be used in a modern educational context, to increase personal reo Māori historical knowledge and to source examples of translations; full text search was the preferred method for finding information because "it's similar to Google"; the language interface was never altered but was always kept in te reo Māori "because I am looking for stuff in Māori"; only extracted text documents were requested as the participant was not interested in the original image; the participant had noticed errors in the extracted texts but was not aware that these were caused by the OCR process; the participant was not aware that quotation marks could be used to assist searching or of the English abstracts and discovered the commentary information only by chance.

UsgMi Interview #3

A research fellow, a history consultant and a translator, aged between 31 and 40 years, this participant was classed as a very heavy user of the Niupepa digital library, accessing it 3–4 hours per day, most working days for four years. The participant had similar experience with other online digital collections. The observation part of the interview lasted 40 minutes, with a further 16 minutes spent responding to the prepared questions.

The participant gave a comprehensive demonstration of using the Niupepa digital library, initially giving two examples of using the search engine to locate documents and altering the Some/All query parameter when necessary. Use of quotation marks to enable phrase searching was observed and examples were given of the Niupepa material being copied and pasted into word processing documents. Sourcing parallel language documents by concurrently entering query terms in te reo Māori and English was demonstrated. Specific documents were shown to be easily located when large blocks of text, copied from other documents, were entered into the query box. Further material about a subject was found by dividing the query term into component parts. For example, a query for ‘Whanganui’ was followed by a query for ‘Whanga nui’ which located more examples. The participant then demonstrated that the English abstracts sometimes contained information that was not located when undertaking a query solely on the newspaper documents.

The following responses were made to the prepared questions: the Niupepa digital library was used to find information on “all topics”, in particular, names, genealogy, traditional areas of residency and reports about what was said at gatherings; full text search was mostly used because “you get there quicker, to the actual particular topic that you are searching for”, however, browsing by date and by periodical listing was also undertaken; the commentary information and English abstracts were used; the participant mostly used reo Māori for the interface and queries, switching the interface language occasionally so that the reference material would be displayed in English and then could be easily copied and inserted into word processing

documents where references were required in English; on other bilingual web sites sometimes the interface language was switched to English "... I haven't got time to learn all those new words, it's laziness really ..."; however, the Niupepa digital library was predominantly used in te reo Māori as the terminology was readily understood.

7.4.2 Interviews With UsgHaw Clients

Five interviews were undertaken with eight users of the Hawaiian Nūpepa digital library. All participants used the digital library with the interface language set to Hawaiian, all had an understanding of written Hawaiian and all were residents in the islands of Hawai'i. Five participants were of Hawaiian ancestry. The interviews were conducted at their homes or places of work; one interview took place on the island of O'ahu and the other four were conducted on the island of Hawai'i. The interviews took place over a two day period in June 2006. All interviews were recorded onto digital video tape, enabling them to be fully documented at a later stage. The participants remain anonymous. All interview recordings on all forms of media are scheduled to be destroyed by December 2007. Brief summaries of the five interviews with usgHaw clients follow.

UsgHaw Interview #1

This participant was a male who did not have Hawaiian ancestry, who worked in the computer industry and was in the over 50 year age bracket. The Nūpepa digital library had been used to assist in his graduate and personal research. He was fluent in both spoken and written Hawaiian, used the Nūpepa digital library perhaps five times or more a week and other online resources perhaps two or three hours everyday. Regarding the interface language he stated, "The truth is, I don't know that I really pay attention to it, when I see the Hawaiian I don't change it, I suppose that if I saw it in English I am not sure that I would not change it either ... I

am happy to see it in Hawaiian.” The observation part of the interview lasted 25 minutes, with a further 20 minutes spent responding to the prepared questions.

The participant demonstrated using the Nūpepa collection to locate a citation by periodical name through the series list. He then clicked on the appropriate date and looked for an extracted text; as there was none he browsed the publication until he found the particular article that he was looking for. He then talked generally about some of the texts he had found in the newspapers and the wealth of literature available there. He demonstrated searching for certain topics and discussed the issues with phrase searching. He spoke about how he had browsed through some earlier newspapers and discovered some of the earliest examples of indigenous language publications by native speakers. He then discussed issues regarding accessing later editions of newspapers.

The following responses were made to the prepared questions: the Nūpepa collection was used to access articles for which he had the citation and to locate new information; the default language of the digital library wouldn't affect usage, “... I am goal driven when I get there ...” but that if the site was suddenly defaulted to the English language he “...would possibly switch back from English to Hawaiian because I suppose that is what I am used to...”; while aware of the purpose of the Some/All query parameter he didn't use it much as he could not “... say that it makes clear sense to me that my searches show up a lot differently for me ...”; the participant would have used the quotation marks (phrase defining query operator) had they been working.

UsgHaw Interview #2

This participant was an information provider, in the over 50 year age bracket and very competent with computers. Of non-Hawaiian ancestry, she had a minimal spoken fluency and medium written fluency in the Hawaiian language. The participant was a heavy user of the Nūpepa collection using it 1–2 hours per day, most days and used other online collections to a similar extent. The observation part of the interview lasted 22 minutes, with a further 15 minutes spent responding to the prepared questions.

The participant was observed undertaking a regular transcription exercise; a Nūpepa document was located by the series list and a transcription of the article was created in a Microsoft Word file. The participant explained how references were obtained from other online and physical sources and then the document was viewed in the Nūpepa web site. The series list was displayed and the publications for each periodical listed. The participant then selected the appropriate time period, requested an article and then began reading the article to locate the topic of interest.

The following responses were made to the prepared questions: the query facility was not used as the participant knew how few of the newspaper documents had been transcribed; query parameters were not altered as the participant was unaware of them and of the detach link^{iv}; usage of the digital library was in Hawaiian because "... I speak about 6 or 7 languages fluently and to switch back and fourth is time consuming. If you can work in the same language you don't have to waste brain time translating".

^{iv} A link that would open the current document view into a new browser window.

UsgHaw Interview #3

The participants were a Hawaiian male and a Hawaiian female. One was a professor in the 31–40 age bracket, the other an editor/researcher/translator in the 21–30 age bracket. Both were fluent in written and spoken Hawaiian. Each accessed the Nūpepa digital library about five times a week and other online resources perhaps 6–12 times per day. The default language of the digital library was always set to Hawaiian although one participant altered this setting to English on the very first visit in order to understand what one particular menu item was referring to. The observation part of the interview lasted 28 minutes, with a further 22 minutes spent responding to the prepared questions.

The participants demonstrated submitting a query for a particular term so they could examine the structure of sentences in which the term was used. They then gave the example of how they had used the Nūpepa collection to find information about an ancestor and stressed how relevant the results had been. One participant explained how students had been required to find articles from the Nūpepa digital library and then to transcribe these into modern Hawaiian. The participants demonstrated the amount of information that could be obtained from transcripts of songs and chants. The participants also showed how they undertook browsing, both on previously referenced articles and on newspapers where particular content or a writer/editor was known.

The following responses were made to the prepared questions: both participants had similar strategies and reasons for accessing the Nūpepa digital library; browsing was preferred when the periodical name and date was known; querying was preferred when it was not known where, or even if, the information existed in the Nūpepa collection "... but then you have the problem that not all the pages are searchable. Am still waiting for that"; neither participants would alter the default language of the digital library, "... I don't see it as being in one language or the other ... I don't really pay attention to the things that I don't need to use ..."; both participants had tried changing the Some/All setting on the query page

but it did not seem to make a difference; quotation marks were used to enable phrase searching, but one participant felt that this did not always seem to work.

UsgHaw Interview #4

Two Hawaiian females participants in this interview, one under 21 years old and one aged over 50 years. One was involved with teacher training, the other a tertiary student and part time writer. Both were fluent in written and spoken Hawaiian. The student used the Nūpepa digital library perhaps 3–5 times a week, the teacher trainer perhaps only twice during weekends. Neither was aware that the language of the interface could be set to anything but Hawaiian or that non-Hawaiian texts (i.e. English) were available in the Nūpepa collection. The observation part of the interview lasted 20 minutes, with a further 24 minutes spent responding to the prepared questions.

The observation section of the interview consisted more of discussion with the interviewer than strict observation of usage, although, the participants did show how they accessed documents. One used other online resources to find citations to the newspapers and then used the Nūpepa series browse and a date identification to select the newspaper required. She then either read the document on the screen or printed it out in a large size so that she could take time to read it. The other mentioned that the image format of the documents was “nice” but as she only used the search facility she mostly accessed the extracted text documents; “... it’s way easier to read and the search terms are highlighted.” Examples were shown of both access methods.

Both participants were surprised that the interface and some articles were available in English. Neither was aware of the query parameters (Some/All) and of the option to detach an article. Both participants stressed how important this resource was for the Hawaiian people.

UsgHaw Interview #5

These two participants were a male (with no Hawaiian ancestry) living in Hawai‘i and a Hawaiian female,^v both in the 31–40 age bracket and fluent in spoken and written Hawaiian. Both participants worked with curriculum development and one was also involved with translation, editing and technical support. One participant was a medium to heavy user of the Nūpepa digital library, using it only once or twice per week but using it for three or four hours at a time and the other participant had not used the Nūpepa collection but had seen it being used. Each was considered a heavy user of other online collections and resources, using them up to 20 times a day. The observation part of the interview lasted 23 minutes, with a further 16 minutes spent responding to the prepared questions.

The Nūpepa digital library was demonstrated primarily by the participant who had used it frequently. Examples were shown of using the digital library to find patterns of word usage and looking up uncommon terms. The participant explained that phrase searching with quotation marks was really useful but did not appear to be working much of the time and mentioned the appropriateness of the icons and accessing extracted text and facsimiles. The participant regularly accessed full size images and used the zoom feature to decipher exactly what some words were or printed out and enlarged the images to assist with deciphering what some characters and words were.

The following responses were made to the prepared questions: these participants were an example of usgHaw clients who preferred to submit queries to find documents but understood that the search facility only searched on a small percentage of the documents and stated “I wish this would get bigger faster”; the Some/All query parameter was not altered, rather the preference was to see all of the query terms all of the time; full size images were preferred despite spending “... twice to three times as long looking at the image ... it’s harder to read but I think it is more accurate. That’s why I do it ... Sometimes I can get led astray by

^v Hawaiian female and Hawaiian male are terms used in this chapter to designate ethnicity not country of residence.

seeing the printed word [OCR extracted text format] and thinking it is accurate"; some of the links were unclear so were not used (for example, mid-size image icon, detach icon) " ...I don't know what they mean so I never use them."

7.4.3 Interviews With UsgEn Clients

Students undertaking computer science degrees at the University of Waikato were interviewed to provide user feedback on the Niupepa digital library from non-indigenous and international perspectives. This particular group was selected because of their availability and the wide variety of nationalities they represented. Although none spoke or read te reo Māori, many of the students were bilingual or multilingual and, consequently were able to provide feedback on their personal experience using electronic resources in a language other than English.

Eight interviews with 19 participants were held over a four day period in March 2006; each interview involved either two or three participants. Although none of the students had used the Niupepa digital library, all had considerable knowledge and experience using online databases. The interviews were conducted in a usability lab where the students' physical actions, verbal interactions and online activity were all videoed simultaneously.

As none of the students had used the Niupepa digital library, the interview format was adapted to include an introduction to the digital library, its purpose, what types of information were available and how they could be accessed. Time was given to the participants to familiarise themselves with the digital library.

The participants remain anonymous. All interview recordings on all forms of media are scheduled to be destroyed by December 2007. Brief summaries of these eight usgEn interviews follow.

UsgEn Interview #1

Both participants were male, aged between 21 and 30 years old. One was from South Africa and the other from Aotearoa. Both were heavy users of online collections, using them about 30 times per week. The interview lasted for 45 minutes; 35 minutes of using the digital library and 10 minutes of answering questions about usage.

After the familiarisation period the participants were asked to locate documents relating to specified topics: documents relating to ‘measles’, to some of their ancestors and to a place called ‘Te Awamutu’. The participants were quick to use the Some/All query parameters and also made use of the quotation marks to undertake phrase querying.

In the final part of the interview the participants were asked questions about their (brief) usage of the Niupepa digital library. They were unclear about the concept and purpose of an English abstract and selected this option for searching as they mistakenly thought a query would return only documents in English. They were not sure how the English abstract texts differed from the texts of the newspaper documents. As these participants could only speak English they were only interested in documents written in English. They suggested that the addition of an icon showing the language of the document in the results and series list pages.

UsgEn Interview #2

One female and one male were interviewed. They were both from China, with Mandarin as their native language and were in the 21–30 age bracket. They were heavy users of online collections, using them about 20–30 times per week. The interview lasted for 47 minutes; 37 minutes of using the digital library and 10 minutes of answering questions about usage.

After the introduction they spent about 10 minutes exploring the interface, reading the home page, the help page and the preferences page. The interface language was switched to Mandarin and a query submitted for the term “Waikato”. The results were examined but they were unsure what some of the numbering in the results pages referred to; when the interface language was switched back to English the numbering scheme became clearer. The participants were asked to find information on measles, a disaster and a famous person. These queries were all undertaken competently.

These participants had some slightly different approaches from those observed in other interviews. Before the activity began they took time to read the about, help and preferences pages where they discovered the option of having the site in Mandarin. The participants remarked that when they were in their own country, China, they preferred to undertake searching activity in Mandarin. Part of the reason they came to Aotearoa was to improve their English and, because assignments had to be written in English, they preferred to undertake this task in English. They also seemed to click through the content very quickly, not pausing very long to read and browse the content of the Niupepa documents.

UsgEn Interview #3

The participants were two male students in the 21–30 age bracket, one from Aotearoa (non-Māori) and the other from Taiwan. Neither was literate in te reo Māori; the Taiwanese participant also spoke Mandarin. Both were heavy users of online collections and online resources, stating they used these over 100 times per week. The interview lasted 35 minutes; 28 minutes of using the digital library and 7 minutes of answering questions about usage.

After the initial introduction, they spent 11 minutes looking around the site and trying out many of the interface controls and browsing options. A very short time was spent looking at each of the interface pages and even less time looking at newspaper documents not written in English. The participants were asked to find three separate pieces of information on diseases, three on natural disasters and

some information on an ancestor, all of which were completed within 14 minutes. These tasks were then followed by the question and answer section of the interview.

Some important points were raised in the interview. These participants preferred to search on the English abstracts as these were “more convenient, if we looked at these we had a better understanding about what it was going on about.” They suggested that in addition to a query term being highlighted in the results page, the page should be automatically scrolled so the query term becomes immediately visible to the user. The bilingual participant mentioned that he only uses a Mandarin interface when he knows that the articles that he is searching for are in Mandarin.

UsgEn Interview #4

The participants were two male students in the 21–30 year old age bracket, one from Zimbabwe who also spoke Afrikaans and the other from China who also spoke Mandarin. Both participants were heavy users of online collections and online resources estimating their use at over 50 times per week. The interview lasted 45 minutes; 35 minutes of using the digital library and 10 minutes of answering questions about usage.

After the introduction they spent 14 minutes exploring the interface, the different browsing options available, submitted some queries and examined the results. They were then asked to find three separate pieces of information about diseases. This task took longer than in some of the other interviews as they spent a lot of time reading the documents. They were then asked to search for information on three disasters and found documents about an earthquake, a flood and a tidal wave. The Mandarin version of the interface was tried and an incorrect translation for volume number was noticed. This concluded the observation section of the interview.

The participants raised some new issues about the usability of the Niupepa digital library. One complained about the (green) colour of the interface and felt that if he had to look at it for too long his eyes would become tired. There was uncertainty whether some of the buttons were clickable. For consistency with other web sites a site logo should be added, the help button should be located on the far right and the home page button on the far left. The Chinese participant explained that usually he uses the Web in Mandarin as it is easier for him to understand. However, if he wanted to search for a sport, such as rugby, then he "cannot search the Chinese [language] website[s] because many Chinese people do not like rugby."

UsgEn Interview #5

The participants were two 21–30 year old male students, one from Germany and the other from Austria. Both also spoke German and were heavy users of online collections and online resources, estimating their use at over 100 times per week. The interview lasted 45 minutes; 32 minutes of using the digital library and 13 minutes of answering questions about usage.

After the introduction they were asked to familiarise themselves with the Niupepa interface. First they set the interface language to German and looked at the different methods for accessing documents and examined some of the documents. They submitted some queries and examined the results. The participants were then asked to find three separate articles relating to diseases. Their initial spelling was poor so they used a German-English online dictionary to get correct spelling, then using the Niupepa query facility found three articles about diseases. The next tasks were to find information on natural disasters and an article about a person. They located articles about three separate earthquakes and about James Cook. Finding these articles completed the observation section.

Some interesting behaviours and responses were recorded in this interview. It was the first where the interface was used for the entire interview in a foreign (non-English, non-Māori) language. The participants only searched the English

abstracts, perhaps because they did not want the results page to be cluttered up with articles that they could not read. They noted that the English abstract page numbering did not always link correctly to the associated newspaper page. Both participants preferred using German when browsing/searching in their free time and English when undertaking course-related work, presumably because their course was taught in English.

UsgEn Interview #6

The participants were three 21–30 year old male German students with English as their second language. All participants described themselves as heavy users of online resources, each accessing these at least 50 times per week. The interview lasted 47 minutes; 32 minutes of using the digital library and 15 minutes of answering questions about usage.

After the introduction they spent 15 minutes familiarising themselves with the interface. The preferences were examined and the interface language was initially switched to German but then switched back to English. They looked at the different browsing methods, accessed several documents, looked at some of the query parameters, submitted a query and then examined some of the query results. The participants were then asked to find articles on three separate illnesses. They entered ‘illness’ into the search engine and searched only on English abstracts. They were surprised that the query term was not found until the interviewer explained that English abstracts only searched on reo Māori documents. The query scope preference was adjusted to ‘papers, commentaries and abstracts’ and then articles were located. Three articles about disasters and an article about a famous person were located without difficulty.

The participants displayed some pertinent usage behaviours. After trying the interface in German they quickly decided to switch it back to English, because they noticed that many sections of the help page had not been translated from English into German. “That’s [in English] how we are used to interact with the computer anyway ... most of the time”. When they first

accessed the series page everything appeared to be in te reo Māori, which was initially daunting and the comment "Let's go back" was made in a joking fashion. The participants made comments about the usability of the interface, the placement of buttons, the labels and what could be added to make the site more usable. They also commented that they rarely undertook language switching of an online database, preferring to use the database in what appears to be the language in which it was built. They also mentioned that if a web site was not easily understandable then they would quickly look for another site.

UsgEn Interview #7

The participants were three 21–30 year old students, one female and two males. They were all from Aotearoa, although one had spent a considerable part of his life in the United States. None was of Māori descent and all described themselves as heavy users of online resources, each accessing online resources between 50 and 100 times per week. The interview lasted 40 minutes; 27 minutes of using the digital library and 13 minutes of answering questions about usage.

After the introduction they were given time to explore the site and looked at the different methods for accessing the documents. Using the date option, they accessed a document and spent time reading the text. They took time to read the help and preferences pages. As one participant had Japanese knowledge they briefly looked at the interface in Japanese before switching it back to English. They returned to the Series option, accessed a publication, looked at the English abstract, and read some background information about the English Abstracts. The participants were then asked to find information about three diseases which they did, although they did have some difficulty trying to spell the diseases correctly. Next they were asked to find information on three natural disasters, which they did without difficulty using the query facility. The final task was to search for a surname which was undertaken without difficulty.

These participants exhibited some unique behaviour. More time was spent reading the Niupepa documents perhaps because the participants were from Aotearoa the

documents had more relevance to them. The series list did not seem to appear so foreign, perhaps because many of the reo Māori titles were familiar to these students. The participants were not particularly interested in the full size images "... if the text was accurate then I wouldn't have any reason to view the full size image ...". They found that they were unable to enter wildcard characters with query terms when undertaking searching. They suggested a spell checker on the search interface would be helpful. There was also the suggestion that pages with highlighted query terms should be automatically scrolled so that the terms were instantly viewable, "generally that's why you go to search pages to go straight to that text".

UsgEn Interview #8

The participants were three students, one female and two males. The female was Austrian and fluent in German and English. She was in the 21–30 age bracket and used online collections or searching perhaps 500 times a week. One of the males was from Aotearoa (non-Māori) in the 41–50 age bracket. He was a part time student, part time stores man who used online resources perhaps 150 times a week. He had no fluency in te reo Māori. The other participant was an Aotearoa Māori with tribal affiliations to Ngāti Porou and Ngāti Raukawa. He had a basic understanding of both spoken and written reo Māori. An educationalist, he accessed online resources perhaps 50 times per week. The interview lasted for 48 minutes; 29 minutes of using the digital library and 19 minutes of answering questions about usage.

After the introduction they were asked to explore the site. They began by looking for information about the Tarawera eruption that occurred in Aotearoa in 1886. Some initial queries returned results but without a preview of the query results the participants were unsure whether the documents were relevant and did not access them. A series browse at first appeared daunting but the titles of the newspapers were explained by the Māori participant and some documents were examined. Two students were happy to see text but one wanted to see the actual image. The

participants undertook some searching on the English abstracts and tried altering the Some/All option. They were asked to find three diseases and to search for some family names, both of which were done without difficulty. In one instance quotation marks had to be used to narrow the results. The family name search concluded the observation portion of the interview.

Some interesting behaviours were observed. When looking at the series list one participant stated, "Problem is none of these mean anything to me, all those results". The Māori participant replied "...this is just giving the title of the newspaper and the volume". Clearly, having some familiarity with te reo Māori improved the usability of this interface in English. It was clear that the participants from Aotearoa were more interested in the texts and wanted to spend longer reading the content. They gave feedback on improving the usability of the interface, for example clearer browsing buttons, a different layout of English abstracts and clearer navigation.

7.4.4 Client Interviews Summary

Clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries were interviewed using two methods. First, the clients were observed using the digital library. Second, the clients were asked questions about why and how they used the digital library. Observations and responses were analysed to determine what factors influenced the clients' use of a particular language.

Three client types were interviewed: those who used the Māori Niupepa digital library with the interface language set to reo Māori (usgMi) all of whom were Māori; those who used the Hawaiian Nūpepa digital library with the interface language set to Hawaiian (usgHaw), both Hawaiian and non-Hawaiian clients; and those who used the Māori Niupepa digital library with the interface language set to English (usgEn), both Māori and non-Māori clients.

The usgMi and usgHaw clients interviewed were experienced users of their particular digital library, all were bilingual and understood the interface in both

the indigenous and non-indigenous languages and were able to read the collection documents in both languages. The usgEn clients interviewed were all experienced users of online electronic resources but none had previous experience with the Māori Niupepa digital library and, aside from one, none were literate in the indigenous (reo Māori) language which meant that they could not use the interface in te reo Māori or comprehend any reo Māori content of the digital library.

With the differences in client types in mind, the factors which affected the use of a particular language in a bilingual digital library environment are summarised in the next three sub-sections.

UsgMi Client Interviews

Only three usgMi clients were interviewed, insufficient to represent all clients who used the Niupepa digital library in te reo Māori. However, these interviews have provided some evidence to suggest factors affecting te reo Māori usage by clients of the Māori Niupepa digital library.

The usgMi participants demonstrated how the Niupepa digital library was used for tribal, personal and academic research. Use of the query facility method was most commonly chosen by the two younger participants, whereas the older participant had a stronger preference for accessing the documents using the series list. One of the younger usgMi participants only accessed the extracted text versions of documents whereas the older usgMi participant mostly accessed the full size images.

The usgMi participant most experienced with the Niupepa digital library, and other online electronic resources, made the most use of query parameters and query operators also using novel ways to improve his querying results, such as using query terms in two languages, separating a single query term into two portions and submitting queries on blocks of texts. The other two usgMi participants made little use of query parameters and query operators.

All usgMi participants showed great interest in the documents located and, consequently, a lot of time was spent reading the documents. They did not mention any negative issues with the Niupepa interface (in te reo Māori), despite two of the participants clearly demonstrating difficulties using the Māori Niupepa interface. One participant had problems with viewing a full size image on the screen, another undertaking queries for phrases. However, neither raised these as problems until they were later queried further on these issues.

All three participants used the interface in te reo Māori because they were searching for Māori topics and felt comfortable in that language. Two mentioned that they used the English version of a bilingual interface if the reo Māori version was not clear.

UsgHaw Client Interviews

Eight usgHaw clients were interviewed. Again, this number is insufficient to represent all clients who used the Nūpepa digital library in Hawaiian, but has provided some evidence suggesting factors affecting Hawaiian language usage by clients of the Hawaiian Nūpepa digital library.

The usgHaw participants demonstrated how the Hawaiian Nūpepa digital library was used for personal research, for education purposes, for Hawaiian language studies and as a source of material for publications. The Nūpepa documents were accessed from browsing by series and by submitting a query. The usgHaw participants generally showed an awareness that only a small proportion of documents had been converted to text and so were prepared to use means other than the query facility to locate documents. In many instances they knew a particular article existed, through references or having previously accessed the article. Consequently, the series list was frequently used to locate an article.

Those with more experience with the digital library tended to view the midsize and full size formats. Those who undertook queries tended to view the extracted

text formats. Querying and use of extracted texts was favoured by the younger participants.

Experienced participants had tried to use the query parameters and query operators but found that they either did not work or did not work as expected and were not used subsequently. Documents relevant or otherwise interesting were often saved to their local computer and/or printed out for more intensive study.

The usgHaw participants did not mention any negative issues with the Hawaiian Nūpepa interface, aside from a common wish to see more documents with extracted text files (thus allowing them to be found by a query). Many stressed how important this resource was for the Hawaiian language and for research into Hawaiian history and culture.

All eight participants used the interface in the Hawaiian language for several reasons; Hawaiian was the default language setting of the Nūpepa interface, as the documents were (virtually all) written in Hawaiian it was cognitively less challenging, and they all expressed a desire to support the use of the Hawaiian language. Some of the participants mentioned (and demonstrated) that they had briefly switched the language of the interface when some terminology used in the Hawaiian language interface was not understood.

UsgEn Client Interviews

Nineteen people were interviewed as clients who used the Māori Niupepa digital library in English. While classified as usgEn clients, they were not representative of all usgEn clients as they were first time users and undertaking set tasks rather than using the digital library through choice. Despite these differences, the usgEn clients interviewed were still able to provide some opinions on using the Niupepa digital library in English and offer some thoughts on the non-English use of online electronic resources.

None of the usgEn participants used the Māori Niupepa digital library in Māori as all but one was unable to understand te reo Māori. The digital library was tried in other languages; one group used it in German for the whole of the interview and other interview groups tried the interface in Mandarin (twice), German and Japanese before switching it back to English.

The usgEn participants were instructed to find information based on certain topics and all used the query facility to achieve this. The series list displayed all the periodical titles but as almost all of these titles were written in te reo Māori (see *Figure 3-4*) this was quite foreign to the usgEn participants. On seeing the series list most chose to return to a page where most of the text was written in English.

The usgEn participants, all experienced users of online electronic resources, readily used the Some/All query parameter and the phrase defining query operator. Many wondered if a spell checker could also be incorporated into the interface. Those who were not from Aotearoa spent little time reading the Māori Niupepa documents while those from Aotearoa often noticed articles of interest to them and subsequently spent a longer time reading documents of the Niupepa digital library.

Many issues about the digital library interface were raised by the usgEn participants: the concept and purpose of the English abstracts was unclear to many of them; they mentioned the query activity would be improved if preview documents were shown in the results page; suggestions were made to improve many aspects of the interface, including where to use more icons, the placement of buttons and links, the colour of the interface and clearer navigation. It seemed the usgEn clients were less interested in the content of the digital library and more interested in critiquing the usability of the interface.

The multilingual usgEn clients mentioned factors which led to their decision to use an electronic resource in a particular language. These included matching the interface language to the content of the resource, matching the interface language to the predominate language of the interface, a desire to use the interface in one particular language (for example, to match the language that the student's course was taught in) and using an interface language that was easily understood.

7.5 Summary

This chapter has discussed two methods used to generate feedback from clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries. Sixty responses were received to a web questionnaire placed on the Niupepa web site in 2006. Sixteen observational studies and interviews were undertaken with clients who used the digital libraries. The web questionnaire responses were too low in number and the clients interviewed too few for the feedback to be representative of all Niupepa/Nūpepa clients. However, the feedback generated has been valuable for two reasons. First, it raised some important usability issues with indigenous languages on bilingual digital libraries. These usability issues are summarised in Chapter 8. Second, the feedback has suggested reasons why indigenous languages are used by clients of digital libraries; specifically why indigenous language speaking clients used these particular digital libraries and why these clients chose to use these digital libraries in the indigenous language.

Feedback from the web questionnaire and client interviews suggests that the Māori Niupepa digital library was used by te reo Māori clients primarily for undertaking tribal research, research on te reo Māori and/or on personal or family information. Research on historic practices, assistance with translations and recreational reading were also reported as reasons that this digital library was used by te reo Māori clients. The range of industries accessing the Māori Niupepa digital library appeared to be mostly limited to academic institutions and tribal organisations with personal researchers also using the resource. The Māori Niupepa digital library appeared to be used in te reo Māori by a small range of industries for a wide range of topics.

Feedback from the client interviews suggests that the Hawaiian Nūpepa digital library has been used extensively in the Hawaiian language education institutions, both at secondary and tertiary levels, often as a prescribed teaching resource, for studies on Hawaiian language, history and culture. In addition the Hawaiian Nūpepa digital library has been used regularly as a source of written material for Hawaiian language publications. The Hawaiian Nūpepa digital library appeared to

be used in the Hawaiian language by a small number of industries for a wide range of topics.

Indigenous language clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries have emphasised how important the digital libraries are because of the content they contain. Clients reported using the content to assist with discussions on history, cultural practices, land tenure and genealogy. Some of the content had even been presented in courts of law. The second aspect was the ‘treasure of the language’ available on the newspaper pages. Indigenous language clients lauded the quality and simplicity of the written indigenous language writings and suggested that making this available assisted in the revival of the traditional indigenous language. Some comments by indigenous language clients include:

“A great source of written information pertaining to the lives of my tupuna [ancestors], iwi [tribe], hapu [sub-tribe] and whanau [family]. In addition, the papers themselves provide interesting accounts of the social and political context of Maori communities at that time. This is incredibly valuable.”

“Kare i tua atu hei patengi kai ma Hinengaro. Ko te mihi ki a Nehe ma i waiho mai e ratou enei tu korero hei whakamaro i toku reo ake kia rite ki ta ratou i korero ai ...”

[There is no greater database to nourish the mind. Salutations to those of the past who left behind these words to strengthen my own (indigenous) language so that it can become similar to the language spoken by them...]

Thus, the content attracted many indigenous language clients to use these two digital libraries. Why did these clients choose to use these electronic resources in the indigenous language?

Feedback from usgMi and usgHaw clients suggested that the indigenous language was used for three primary reasons:

- the content was mostly in the indigenous language and it was cognitively less challenging to have the interface also in the indigenous language
- as the default language of the interface was normally set to the indigenous language, clients using the digital library would often not notice which interface language it was set to
- clients wanted to support the use of the indigenous language. They stated they want to use the indigenous language in as many avenues as possible and digital libraries provide another avenue to do this.

Indigenous language clients also mentioned that if the indigenous language version of the interface was not clear they would switch to the English language version to understand what was being displayed. Once the interface was understood it was then generally switched back to the indigenous language for the reasons listed above.

Feedback from the multilingual usgEn participants also supported this view, suggesting that they also used an interface in a particular language because of the content, the default language setting and the desire to use one particular language. This feedback suggests that content, default language setting and desire to support indigenous language use affected all multilingual clients (both indigenous and non-indigenous) when making decisions about which language to use in a multilingual online interface.

This concludes the chapter about why indigenous languages were used by clients of digital libraries. The final chapter summarises the findings of this thesis and the thesis statement will be revisited.

8 Research Findings

The research described in this thesis has examined indigenous language usage in a digital library environment that has been accessed via the Internet.

Using the Māori Niupepa and Hawaiian Nūpepa digital libraries, this research has shown how indigenous languages were used by choice in these electronic environments during 2005. The results provide encouragement and optimism to people who are striving to retain, revitalise and increase the use of indigenous languages in information technologies. The Transaction Log Analysis (TLA) methods used in this research serve as an example of how web logs can be used to provide significant information about language usage in bilingual online information systems. Combining the TLA with user feedback has provided insights into how and why clients use indigenous languages in their information retrieval activities. These insights in turn, show good practise that is relevant not only to those working with indigenous languages, indigenous peoples or multilingual environments, but to all information technology designers who strive for universal usability.

This thesis comprised eight chapters. The first chapter described why it is important to investigate the use of indigenous languages in electronic environments and proposed that digital libraries can assist in the usage, and subsequent revitalisation, of these languages. The second chapter introduced TLA and explained how this methodology was used to generate data about language usage. This methodology was then used in the next four chapters to analyse: te reo Māori usage in the Niupepa digital library (Chapter 3); international usage of te reo Māori in the Niupepa digital library (Chapter 4); the effect of the default

language setting on te reo Māori usage (Chapter 5) and on Hawaiian language usage of the Nūpepa digital library (Chapter 6). Some user feedback was described in Chapter 7, which suggested reasons why clients used the indigenous language in digital libraries. Finally, Chapter 8 summarises the key findings from the previous chapters and returns to the thesis claim that digital libraries can be used to assist in the usage of indigenous languages.

8.1 Reo Māori Usage

Te reo Māori was used in the digital library environment.

Results from TLA show that te reo Māori was used in 37,936 requests to the Niupepa digital library in 2005 an average of over 100 requests in te reo Māori per day. Te reo Māori requests to the Niupepa digital library comprised 20.5% of the total requests.

It is difficult to accurately calculate the percentage of potential Internet clients who have an ability to undertake online activities in te reo Māori. A 2001 study indicated that only 35% of Māori (in Aotearoa) had access to the Internet compared with 51% of New Zealand Europeans (Te Puni Kōkiri, 2001a). The 2006 census indicated that only 24.4% of Māori were literate in te reo Māori (Statistics New Zealand, 2006). Applying, with some reservations, the 2001 access figures to the 2006 census numbers:ⁱ

- 35% * 565,329 = 197,865 Māori with Internet access
- 24.4% * 197,865 = 48,279 Māori speaking people with Internet access
- 65% * 3,462,618 = 2,250,701 non-Māori people with Internet Access
- 2,298,980 of population in Aotearoa have Internet Access
- 2% of the Aotearoa Internet Access population speak te reo Māori.

ⁱ These figures must be treated with caution. No allowance has been made for the increase in Internet users from 2001 to 2006, for the predominant age group of Māori speakers and online users, for the type of Internet access, or for the non-Māori speakers of te reo Māori.

The conservative estimate calculated above suggests that as little as two percent of the population in Aotearoa have the ability to access and also understand online information in te reo Māori. Given this low percentage, and without considering other factors such as who would want to access the Niupepa, the fact that over 20% of Niupepa requests were with a reo Māori language setting suggests that this digital library supports, and perhaps, encourages, usage in te reo Māori.

User feedback has suggested that the content of the web site, the default interface language and the desire to use the web site in te reo Māori have been factors which led to use in te reo Māori.

8.2 International Reo Māori Usage

Te reo Māori was used by non-Aotearoa clients.

Results from TLA showed that approximately 10% of requests to the Niupepa digital library were requests from international (non-Aotearoa) clients. International clients were almost three times more likely to use the interface in te reo Māori than national clients. Of the 17,390 international requests, 45.4% were with the interface language set to reo Māori while of the 155,515 national requests, 16.5% had the interface language set to reo Māori. International clients were most likely to use the interface in te reo Māori when the default language of the web site was set to reo Māori.

International clients who used the Niupepa digital library in te reo Māori exhibited different behaviours from national clients who used te reo Māori. International clients appeared to be experienced in information retrieval and confident to use the interface in English or te reo Māori. The use of query parameters and query operators and the apparent accuracy of the query activities indicated that these clients were familiar with information retrieval activities. A comparatively higher proportion of requests for parallel (reo Māori/English) documents and English language documents suggested that international clients were seeking information for English language purposes more often than national clients.

Use of te reo Māori by international clients in the Niupepa digital library is an example of how an electronic resource such as the Niupepa digital library can be used to assist in the international usage of an indigenous language.

8.3 Default Language Effect

Setting the default language of the interface to an indigenous language increased clients' use of the indigenous language.

Analysis has shown that there were 21% more reo Māori usage requests when the default language was set to te reo Māori. These requests were made by approximately the same number of clients, indicating that setting the default language to reo Māori did not encourage new clients to use te reo Māori, rather it encouraged bilingual clients to use te reo Māori more often.

When the default language was set to te reo Māori the reo Māori usage sessions were longer with more requests. These sessions also contained more browsing by series and more requests for the extracted text formats of documents. In addition, when the default language was set to te reo Māori, clients classified as reo Māori clients requested significantly more documents written in te reo Māori and significantly fewer documents written in parallel text (reo Māori and English). These clients also submitted more queries, more query terms per query and used a higher proportion of te reo Māori query terms.

These results indicate that setting the default language of a bilingual interface to an indigenous language not only results in an increase in use of the indigenous language but also promotes more situations where the indigenous language can be used. Clients commented that it was good that the indigenous language was the default language; they quickly became accustomed to it and would switch back to it if they found the indigenous language was not the default language.

8.4 Hawaiian Language Usage

The Hawaiian language was used in the digital library environment.

Results from TLA undertaken on the Hawaiian Nūpepa digital library showed that the Hawaiian language was used in 112,786 requests in 2005, 53.4% of the total (filtered) Hawaiian Nūpepa requests. When the requests were classified into language of usage categories, approximately half were Hawaiian usage requests, one quarter were English usage requests and the remaining quarter were classified as bilingual usage requests.

Feedback from Hawaiian language users, similar to the reo Māori feedback, has suggested that this digital library was used in the Hawaiian language because it was the default language of the interface, it matched the content of the digital library and the users expressed a desire to use the Hawaiian language in as many avenues as possible.

While many of the information retrieval behaviours of Hawaiian language clients matched those of reo Māori clients there were some notable differences. There were more indigenous language requests to the Hawaiian Nūpepa digital library than to the Māori Niupepa digital library. One reason for this difference could be attributed to the content of the respective digital libraries. The Hawaiian Niupepa digital library content was primarily in the Hawaiian language with an insignificant amount available that was written in English. Consequently, this digital library has little value to clients who are not literate in the Hawaiian language. The Māori Niupepa digital library has 46% of its content written in parallel text (reo Māori and English) and a further 23% contains English abstracts explaining what the reo Māori content is about. Consequently, over two thirds of the Niupepa content has value to clients who are not literate in te reo Māori. As many more potential clients who were monolingual (English literate) would be able to understand much of the content it seems likely that this larger group of clients would submit the most number of requests.

8.5 Indigenous and non-Indigenous Usage

Indigenous language clients exhibited different information retrieval behaviours from non-indigenous language clients

TLA revealed that the information retrieval preferences of indigenous language clients were different from the preferences of non-indigenous language clients. While both types of clients predominantly used queries to locate documents, the non-indigenous language clients had a higher preference for this method. The indigenous language clients had a stronger preference for locating documents by the series list. User feedback suggested that this was because the indigenous language clients often had prior knowledge of which publication series they were interested in. As the publications were mostly indigenous language titles, the series list was mostly written in the indigenous language, which tended to alienate the non-indigenous language clients. The non-indigenous language clients also had a higher propensity for accessing documents by using the date facility, whereas the indigenous language clients tended not to use this method. This difference could be attributed to different values and different attitudes that cultures display towards time (Hall, 1983).

Indigenous language clients preferred to access documents in the indigenous language while non-indigenous (English) language clients preferred to access parallel language documents and English language documents. Statistics from TLA highlighted this behaviour and it was not unexpected.

Queries undertaken by indigenous language clients often involved a greater number of query terms. It can be expected that clients are likely to enter more query terms in a language they are familiar with. A higher number of query terms entered by indigenous language clients meant a correspondingly greater use of the Some/All query parameter and of the quotation marks phrase matching operator. Query terms submitted by indigenous language clients were mostly in the

indigenous language, matching the language of the content, and appeared to be more successful than non-indigenous language query terms overall.

International (non-Aotearoa) clients exhibited different information retrieval behaviours from national (Aotearoa) clients. Further, the behavioural differences noted between international and national reo Māori clients were not the same as the behavioural differences noted between international and national English language clients. These observations suggest that the international indigenous language client uses unique information retrieval behaviours, behaviours that are unlike the national indigenous language retrieval behaviours and the international non-indigenous language behaviours.

The default language setting of the digital library affected indigenous and non-indigenous language clients differently. When the default language was set to the indigenous language, usage requests in that language increased and when the default language was set to the non-indigenous language, usage requests in the non-indigenous language increased. These results were not surprising but what was unexpected was the results that occurred when the interface language did not match the client's preferred interface language. When the default interface language was set to the indigenous language there was a reduction in the number of usage requests by non-indigenous language clients and the non-indigenous language clients reduced by a third. When the default language of the interface was set to the non-indigenous language there was a similar reduction in the number of usage requests in the indigenous language but no reduction in the number of indigenous language clients. These characteristics suggest that the indigenous language client was a bilingual client who could tolerate the default language of the interface being in either language. However, the non-indigenous language client, perhaps not being bilingual, was often intolerant of the default language not matching their preferred interface language and, therefore, a third chose not to use the digital library in this situation.

8.6 Bilingual Users

The interface was often used bilingually to enhance the information retrieval activity.

There appeared to be a deliberate strategy by some clients of the Māori Niupepa and Hawaiian Nūpepa digital libraries to use the interface bilingually to enhance information retrieval. These clients submitted the greatest number of queries per session (though by definition these sessions were long), altered the Some/All query parameter the most frequently and submitted the greatest mix of indigenous and non-indigenous language query terms. Queries submitted by these clients appeared to be the most effective in locating the required documents.

The information retrieval behaviours displayed by bilingual clients of the Māori Niupepa digital library were very similar to those by bilingual clients of the Hawaiian Nūpepa digital library. What was not similar was the number of bilingual usage requests made to each of the digital libraries; 3,921 bilingual usage requests were made to the Māori Niupepa digital library which represents 3.2% of the total Niupepa usage requests and 50,359 bilingual usage requests to the Hawaiian Nūpepa digital library representing 25.4% of the total Nūpepa usage requests.

The design of the interface was perhaps the primary reason for significantly more bilingual usage requests to the Hawaiian Nūpepa digital library than to the Māori Niupepa digital library. The Hawaiian Nūpepa interface allowed clients to toggle the interface language between Hawaiian and English on each interface page while the Māori Niupepa interface only offered this toggle facility on the home page. Māori Niupepa clients wanting to change the interface language had to return to the home page or access the preferences page. This extra action appears to have been too much trouble for many of the potential bilingual clients because the Māori Niupepa interface was rarely used bilingually.

8.7 Interface Usability and Design

The TLA and user feedback showed some usability issues with the interface design of both digital libraries. Most of the critiquing appeared to be from non-indigenous clients; indigenous clients appeared less concerned with usability issues and/or more reluctant to mention problems which they had encountered. There are three potential reasons why indigenous clients appeared unwilling to mention interface problems: because indigenous clients often had more interest in the texts which led them to disregard the flaws in the interface; because indigenous clients were less experienced in online interfaces and, consequently, did not have high expectations; or perhaps the unwillingness is part of the indigenous culture where belittling a person or their work is not seen as constructive. Whatever the reasons were, less negative feedback was noted from indigenous language clients.

Although the usability evaluation of these digital libraries was not intended as part of this research, it is apparent that if certain usability issues were addressed then the use of the indigenous language is likely to increase. Some of the usability issues are detailed in this section to assist other interface designers intending to design bilingual interfaces and/or interfaces that use indigenous languages.

Translations

Translations should be unambiguous, accurate and relevant to the context. The te reo Māori translation for the link “About English Abstracts” was listed as “He Rāpopototanga Reo Pākehā”. This phrase could be misconstrued to mean “a summary in English”. TLA indicated the reo Māori version of this link was requested at an unusually high rate, which suggests confusion and, thus, ambiguity. Some of the Chinese users also noted some confusion in the Chinese translations of newspaper chapter numbers and, subsequently, chose not to use the

interface in Chinese. These examples show that inaccurate translations can influence language usage in digital libraries.

The site should be translated completely. Some multilingual users mentioned that if they noted a site partially translated they would switch the interface language to the dominant language.

Bilingual sites do not need direct translations. TLA indicated that clients who used the site in the indigenous language were most interested in indigenous language documents. Clients who used the site in the non-indigenous language were most interested in the non-indigenous language documents. Consequently, an indigenous language version of the interface should focus on highlighting and directing clients to the indigenous language content in the first instance. In a similar manner, perhaps a non-indigenous language version of the interface should highlight and direct clients to the non-indigenous language content in the first instance.

Content Description and Format

The different types of content should be clearly described. Observation of clients using the Māori Niupepa digital library indicated that they were often unsure if they were reading newspaper documents or English abstracts about the newspaper documents. They were also unaware that the English abstracts only summarised te reo Māori documents. Users were observed searching the English abstracts for English language texts, only to be disappointed when the results returned only links to te reo Māori documents.

The abstracted text documents need disclaimers. Many clients were unaware that the extracted text documents had been generated by an OCR process that was not 100% accurate. While errors in the extracted texts of documents were sometimes noted and commented on by clients, many clients were unaware that the full size image of the document was an exact replica and, thus, contained no extraction errors.

Image files stored in pdf format with a high resolution can easily be enlarged to assist with reading. The pdf format also allowed the document pages to be easily printed.

A list of titles predominantly in one language should also include translations for the benefit of clients who do not understand that language. Even with the interface set to English, the series list (see *Figures 3-4* and *6-4*) was perceived as foreign by clients who did not understand the indigenous language. Parallel translations of the titles could assist those clients.

Query Activity

The number of documents that are searchable should be made clear. Many clients assumed that all digital library documents were searched when a query was submitted. However, 98.6% were searched in the Māori Niupepa digital library and only 6.4% in the Hawaiian Nūpepa digital library.

The more searchable documents the better. Both TLA and the user interviews indicated that an important feature of the digital libraries was the ability to undertake full text searching. Clients of the Hawaiian Nūpepa digital library expressed an eagerness for more document pages to become available for full text searching as quickly as possible.

Query parameters and query operators should function correctly. The query parameters and query operators appeared to act incorrectly or inconsistently in one of the digital libraries. Consequently, clients could not rely on these to work correctly and use of the operators was subsequently abandoned.

Search results should contain preview images of the returned documents. A preview image of the result, especially when text surrounding the query term was included, allowed clients to decide quickly if the document was relevant to the query. Results could be scanned much more quickly and the search activity was more efficient when preview documents were available. Clients also suggested

that when a document was displayed following a query, the document should be automatically scrolled to the query term position and the query term highlighted.

The user interviews revealed how surprisingly poor digital library users were at spelling. A spell checking function incorporated into the query term input box would improve the submission of queries. If the spell checker worked with both the indigenous and non-indigenous languages it would be even more useful.

Navigation and Layout

Navigation should be clear and logical. Users commented that at times they were unsure where they were in the digital library and how they could return to a previous screen. Buttons and icons should afford their purpose. User feedback suggested that clients were not aware of the purpose of some buttons and icons simply by looking at them and that many were unaware that some buttons were, in fact, buttons; they were written as text with no sign that they were clickable hyperlinks.

The screen layout should be consistent with common practice. Users mentioned that navigation would be easier if the buttons were located in a similar position to that used on other web sites; the home button at the top left and the help button at the top right. The button locations should be consistent throughout the site.

A language toggle button on every page of the interface would assist clients to interact with the digital library bilingually.

8.8 Results Requiring Further Investigation

The Transaction Log Analysis used in Chapters 3-6 yielded a large amount of information. While all relates to indigenous language usage in bilingual digital libraries and, as such, is relevant to this research, much of this information is

beyond the scope of this thesis. Further work is required in many research areas to complete the analysis of indigenous language usage in digital libraries.

For example, an early section in Chapters 3-6 discussed results from how clients explored but then did not go on to use a digital library. Data was provided about the lengths of these exploratory sessions, the apparent read times of interface pages and which interface pages clients requested before deciding not to go on and use the digital library. This data is relevant to investigations into why certain online resources/web pages are used while others are just briefly explored. The differences shown in indigenous language and non-indigenous language exploratory sessions would give insights into online exploratory differences between indigenous and non-indigenous people. Further investigation of these results is required by researchers in the field of human-computer interaction (HCI).

Another topic requiring further investigation is the propensity for indigenous language clients to request documents published in certain years (see *Figures 3-23* and *6-16*). Why were these particular years favoured? Because of historical events that happen in those years, because those were the years that most of the periodicals were published, because those were the documents with the most extracted text formats or because certain authors were publishing certain content in those years? Further investigation is required by historians, sociologists and even librarians to find answers to these questions.

The data generated about interface language switching (see §3.4.7, §4.3.5, §5.3.5 and §6.4.6) also requires further investigation. It appears that indigenous language clients alter the language of the interface a lot less often than non-indigenous language clients. This could suggest that many of the indigenous language clients are bilingual or multilingual. The point at which the interface language is switched within the sessions could suggest a frustration by clients with that particular interface. A consistent level of switching by a user group through a session could indicate a deliberate strategy of using the interface bilingually. Applied linguistics research could provide answers about why these switching behaviours occurred.

The research described in this thesis examines activity in the year 2005. The same methods could be used to examine activity recorded in 2004, 2006, and in subsequent years. A longitudinal study of this nature would provide valuable information about how indigenous language use was evolving in online information systems.

8.9 Claims Revisited

In the introduction the following claim was made:

This thesis charts indigenous language usage by clients of a digital library and demonstrates how digital libraries can provide an environment to support the use of indigenous languages.

Subsequent chapters used Transaction Log Analysis (TLA) to chart indigenous language usage in the Māori Niupepa and Hawaiian Nūpepa digital libraries. The TLA has shown how the indigenous language was used by clients in these digital libraries and some settings that appear to encourage indigenous language usage. Feedback from digital library clients has proposed why the indigenous language was used and has suggested some design principles that would further promote indigenous language usage.

This thesis has shown that indigenous languages were used in digital library environments in 2005. The indigenous language use was discretionary; that is, it took place without coercion. Indigenous languages were successfully used to retrieve information in a modern information system. Using te reo Māori and the Hawaiian languages as examples, this research has shown how digital libraries can provide an environment to support the use of indigenous languages.

8.10 Final Comments

This thesis began with reo Māori poetry from a whakawātea that emphasised the importance of the te reo Māori. It is appropriate to also end with reo Māori poetry, this time taken from a hakaⁱⁱ entitled, *Tōku Reo! Tōku Reo!* ('My Language! My Language!')ⁱⁱⁱ The words are an encouragement and a challenge to Māori people to take responsibility for te reo Māori survival into the future. The challenge has been laid, it is up to Māori to seize this challenge and assist with the revitalisation of te reo Māori. No translation is given.

E tātou e te ao Māori, taringa whakarongo!
Ko te reo te moko o te hinengaro!
Ki te mate tō tātou reo
Ehara kē i te mea me mate
He kore kē ia nō tātou
I aro atu, i hautoa kia ora tonu ai
Nei rā te mātuka kua takoto
Kawe ake! Kawe ake!
Hei!

ⁱⁱ haka: a Māori dance with postures.

ⁱⁱⁱ This haka was composed by Tīmoti Kāretu and first performed in 1994 by the Te Whare Wānanga o Waikato kapahaka (performing arts) group at national competitions in Te Hāwera.

References

- Alosius, J. (2007). Students Enrolled in Public Provider by Ethnic Group in 2005 (Spreadsheet Provided by Ministry of Education). Wellington: Data Management Unit, Ministry of Education.
- Apperley, M. D., Keegan, T. T., Cunningham, S. J., & Witten, I. H. (2002). Delivering The Māori Newspapers on the Internet. In J. Curnow, N. Hopa & J. McRae (Eds.), *Rere Atu Taku Manu! Discovering History Language and Politics In The Māori Language Newspapers*. (pp. 211–236). Auckland: Auckland University Press.
- Barbour, R. H., Cunningham, S. J., & Ford, G. (1993). Maori Wordprocessor for Indigenous New Zealand Young Children. *British Journal of Educational Technology*, 24(2), 114–124.
- Bedford, R., Didham, R., Ho, E., & Hugo, G. (2004). Maori Internal and International Migration at the Turn of the Century: An Australasian Perspective. *New Zealand Population Review*, 30(1/2), 131–142.
- Benton, R. A. (1978). *Can the Maori Language Survive?* Wellington: New Zealand Council for Educational Research.
- Benton, R. A. (1981). *The Flight of The Amokura*. Wellington: New Zealand Council for Educational Research.
- Benton, R. A. (1996). Making the Medium the Message: Using an Electronic Bulletin Board System for Promoting and Revitalising Maori. In M. Warschauer (Ed.), *Telecollaboration in Foreign Language Learning* (pp. 187–204). Honolulu, HI: Second Language Teaching & Curriculum Center, University of Hawai'i.
- Benton, R. A. (1997). *The Māori Language: Dying or Reviving?* Wellington: New Zealand Council for Educational Research.

- Blecic, D., Bangalore, N., Dorsch, J., Hendersen, L., Koenig, M., & Wellar, A. (1998). Using Transaction Log Analysis to Improve OPAC Retrieval Results. *College and Research Libraries*, 59(1), 39–50.
- Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking Questions: The Definitive Guide to Questionnaire Design-For Market Research, Political Polls, and Social and Health Questionnaires*. San Francisco: Jossey-Bass.
- Buszard-Welcher, L. (2001). Can the Web Help Save My Language. In L. Hinton & K. Hale (Eds.), *The Green Book of Language Revitalisation in Practise* (pp. 331–345). San Diego, California: Academic Press.
- Cartwright, J., Chantiny, M., Hori, J., & Peacock, K. (2000). The Digital Landscape: Hawiian Newspapers, War Records, and Trust Territory Image Repository of the University of Hawaii [Electronic Version]. *First Monday*, 5, 15. Retrieved 26 June 2006 from:
www.firstmonday.org/issues/issue5_6/cartwright/index.html.
- Chapin, H. G. (1996). *Shaping history the role of newspapers in Hawai'i*. Honolulu: University of Hawai'i Press.
- Chau, M., Fang, X., & Sheng, O. (2005). Analysis of the Query Logs of a Web Site Search Engine. *Journal of the American Society for Information Science and Technology*, 56(13), 1363–1376.
- Crystal, D. (2000). *Language Death*. Cambridge: Cambridge University Press.
- Cunliffe, D. (2004). Promoting Minority Language Use on Bilingual Websites. *Mercator Media Forum*, 7(1), 42–53.
- Dallimore, G. (1990). He Arahi, He Tohu o Nga Pepa Maori: A bibliography of Maori Newspapers, 1840–1900. Unpublished manuscript.
- Durie, M. (2004). Understanding Health and Illness: Research at the Interface Between Science and Indigenous Knowledge. *International Journal of Epidemiology*, 33(5), 1138–1143.
- Dyson, L. E., Hendicks, M., & Grant, S. (2006). *Information Technology and Indigenous People*. Hershey PA: Information Science Publishing.
- Fishman, J. A. (1991). *Reversing Language Shift: Theoretical and Empirical Foundations of Assistance to Threatened Languages*. Clevedon: Multilingual Matters.
- Gordon, R. G. (2005). *Ethnologue: Languages of the World*, Fifteenth edition. Electronic Version. Retrieved January 2007 from: www.ethnologue.com.
- Grenoble, L. A., & Whaley, L. J. (2006). *Saving Languages. An Introduction to Language Revitalization*. New York: Cambridge University Press.

- Hall, E. T. (1983). *The Dance of Life: The Other Dimension of Time*. New York: Anchor Press/Doubleday.
- Harlow, R. (1987). *A Word-List of South Island Maori*. University of Auckland: Linguistic Society of New Zealand.
- Jansen, B. (2006). Search log analysis: What it is, what's been done, how to do it. *Library and Information Science Research*, 28(3), 407–432.
- Jansen, B., & Spink, A. (2005). How Are We Searching the World Wide Web? A Comparison of Nine Search Engine Transaction Logs. *Information Processing and Management*, 42, 248–263.
- Jansen, B., Spink, A., & Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries On the Web. *Information Processing and Management*, 36(2), 207–227.
- Jones, S., Cunningham, S. J., McNab, R., & Boddie, S. J. (2000). A Transaction Log Analysis of a Digital Library. *International Journal on Digital Libraries*, 3(2), 152–169.
- Keegan, P. (1996, 14-17 January). *Initiatives in Maori Networking*. Proceedings of the 6th Biennial Conference of the New Zealand Computers in Education Society, Hamilton, New Zealand (pp. 35–52).
- Keegan, T. T., Apperley, M., Cunningham, S. J., & Witten, I. (2001, 23-26 July). *The Niupepa Collection: Opening the Blinds on a Window to the Past*. Proceedings of the ICHIM01 International Cultural Heritage Informatics Meeting Milan, Italy (pp.347–356).
- Keegan, T. T., & Cunningham, S. J. (2005a, 7–11 June). *Language Preference in a Bi-language Digital Library*. Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, Denver, Colorado, USA. (pp. 174–175).
- Keegan, T. T., & Cunningham, S. J. (2005b, May 26–28, 2005). *What Happens If We Switch The Default Language of a Website*. Proceedings of the First International Conference on Web Information Systems and Technologies, Miami, Florida (pp. 263–269).
- Kāretu, T. (1996). *Ngā Waiata me ngā Haka a te Kapahaka o Te Whare Wānanga o Waikato (He Whakahoutanga)*. Hamilton: University of Waikato.
- Laws, M. R. (1995). *A Computer Assisted Māori Language Laboratory. The Design, Development and Implementation of an integrated package of Māori language course material in a computerised environment—is it*

- possible and practical*. Proceedings of the Inaugural NAMMSAT Conference, Auckland University.
- Laws, M. R. (2001). *Māori Language Integration in the Age of Information Technology: A Computational Approach*. Unpublished PhD Thesis, University of Otago, Dunedin.
- Metge, J. (1967). *The Maoris of New Zealand*. London: Routledge & K. Paul.
- Nichols, D. M., Witten, I. H., Keegan, T. T., Bainbridge, D., & Dewsnip, M. (2005). Digital Libraries and Minority Languages. *New Review of Hypermedia and Multimedia*, 11(2), 139-155.
- Nielsen NetRatings. (2005). Reaching NZ Māori Online. Retrieved January 2007 from: www.nielsen-netratings.com/pr/pr_050420_newzealand.pdf.
- Office of the Domain Name Commissioner. (2005). Second Level Domains Electronic Version. Retrieved April 2006 from: www.dnc.org.nz.
- Peters, T. A. (1993). The History and Development of Transaction Log Analysis. *Library Hi Tech*, 11(2), 41-66.
- Pool, I. (1991). *Te Iwi Māori: A New Zealand Population Past, Present & Projected*. Auckland, New Zealand: Auckland University Press.
- Reedy, T. M. (2000). Te Reo Māori: The Past 20 Years and Looking Forward. *Oceanic Linguistics*, 39(1), 157-169.
- Russell Bernard, H. (1992). Preserving Language: Computers Can Be a Tool For Making the Survival of Languages Possible. *Cultural Survival Quarterly*, 16(3), 15-18.
- Silverstein, C., Henzinger, M., Marais, H., & Moricz, M. (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33(1), 6-12.
- Smith, L. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. Dunedin: University of Otago Press.
- Stauffer, R. (2006). Personal communication, June 13, 2006, Honolulu, Hawai'i.
- Statistics New Zealand. (2006). 2006 Census Data. Retrieved January 2007 from: www.stats.govt.nz/census/2006-census-data/.
- Tan, P., & Kumar, V. (2002). Discovery of Web Robot Sessions Based on Their Navigational Patterns. *Data Mining and Knowledge Discovery*, 6(1), 9-35.

- Te Puni Kōkiri. (1998). *The National Māori Language Survey: Te Mahi Rangahau Reo Māori*. Wellington: Te Puni Kōkiri (Ministry of Māori Development).
- Te Puni Kōkiri. (2001a). Māori Access to Information Technology. Retrieved January 2007 from: www.tpk.govt.nz/publications/.
- Te Puni Kōkiri. (2001b). Te Māori i Ngā Rohe. Retrieved April 2006 from: www.tpk.govt.nz/publications/.
- Te Puni Kōkiri. (2004). Te Reo Māori i te Hapori—Māori Language in the Community [Electronic Version]. Retrieved January 2007 from: www.tpk.govt.nz/publications/.
- Walker, R. (2005). *Quality Assurance in Tertiary Education from a Māori (Indigenous) Perspective*. Proceedings of the International Network for Quality Assurance Agencies in Higher Education (INQAAHE), Wellington, New Zealand.
- Wang, P., Berry, M., & Yang, Y. (2003). Mining Longitudinal Web Queries: Trends and Patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743–758.
- Warschauer, M. (2000). Language Identity and the Internet. In B. Kolko, L. Nakamura & G. Rodman (Eds.), *Race in Cyberspace* (pp. 151-170). New York: Routledge.
- Warschauer, M., & Donaghy, K. (1997). Leokī: A Powerful Voice of Hawaiian Language Revitalization. *Computer Assisted Language Learning*, 10(4), 349–361.
- Williams, H. W. (1985). *A Dictionary of the Maori Language* (Seventh ed.). Wellington: Government Printing Office.
- Wilson, W. H. (2006). Personal Communication. In T. T. Keegan (Ed.). Hilo: Ka Haka 'Ula O Ke'elikolani College of Hawaiian Language, University of Hawai'i at Hilo.
- Witten, I. H., & Bainbridge, D. (2002). *How to Build a Digital Library*. San Francisco, CA: Morgan Kaufmann.

Appendix A Pearl Scripts Summary

Perl scripts were written to process the raw web logs. A description of their use in generating Transaction Log Analysis data is given in §2.5. The Perl scripts filtered the web logs to remove inappropriate requests, classified the requests into sessions and generated statistical information. The output of the Perl scripts were text files which were entered into spreadsheets so that graphs and further statistical information could be generated. The Perl scripts were written by Gareth Judsen under supervision in 2005/2006 while Gareth was a student of the Computer Science Department at the University of Waikato. Summaries of the individual Perl scripts are tabled below.

Many of the Perl scripts require a specific directory structure to run correctly. This structure consists of:

<base directory>—directory containing the initial raw log and may have any name
<base directory>/Results—directory to receive the output text files
<base directory>/Results/Reading Times—directory to receive the reading files
<base directory>/Session Data—directory to receive the sessions files

A series of tables below describe each of the Perl scripts, and lists the input files required and the output files produced by the Perl scripts. This is followed by a list of data and statistics produced by the Perl scripts.

To obtain a copy of the Perl scripts contact the author, Te Taka Keegan at the Computer Science Department, University of Waikato, Aotearoa.

Name:	process-raw-logs.pl
Description:	Used for processing raw log to remove unwanted CGI arguments and unwanted requests, such as requests to incorrect collection, non-English and reo Māori requests, requests from web robots, etc. Outputs summary information indicating how many requests were removed and why, lists web robots found by agent string in robot_agents.txt and web robots found by IP address in robot_ips.txt.
Size of Script:	319 lines of code
Input Files:	Raw log (can have any name) e.g. <base directory>/RawLog.txt
Output Files:	<base directory>/PROCESSED-RawLog.txt (i.e. the input file with the string 'PROCESSED-' added to it)
Additional Requirements:	The script takes the input file and year for the requests as a command line argument, in that order. The script requires 'logs.pl' 'robot_agents.txt' and 'robot_ips.txt'; all should be located in the same directory as this script.
<i>Table A-1: Summary of script process_raw_logs.pl.</i>	

Name:	create-sessions.pl
Description:	Creates session files from the processed logs. Produces a results file summarising the counts of the various sessions.
Size of Script:	387 lines of code
Input Files:	<base directory>/processedRawLog.txt
Output Files:	<base directory>/Session Data/Zargument-singles.txt <base directory>/Session Data/Zargument-exploratory.txt <base directory>/Session Data/Zargument-usage.txt <base directory>/Session Data/IP-singles.txt <base directory>/Session Data/IP-exploratory.txt <base directory>/Session Data/IP-usage.txt <base directory>/Results/General Session Data.txt
Additional Requirements:	The script takes the input file as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.
<i>Table A-2: Summary of create-sessions.pl.</i>	

Name:	split-sess-lang.pl
Description:	Divides the session files into languages of session files. Generates switching data and files.
Size of Script:	509 lines of code
Input Files:	<base directory>/Session Data/Zargument-exploratory.txt <base directory>/Session Data/Zargument-usage.txt <base directory>/Session Data/IP-exploratory.txt <base directory>/Session Data/IP-usage.txt
Output Files:	<base directory>/Session Data/Zargument-exploratory-Maori.txt <base directory>/Session Data/Zargument-exploratory-English.txt <base directory>/Session Data/Zargument-usage-Maori.txt <base directory>/Session Data/Zargument-usage-English.txt <base directory>/Session Data/Zargument-usage-Bilingual.txt <base directory>/Session Data/IP-exploratory-Maori.txt <base directory>/Session Data/IP-exploratory-English.txt <base directory>/Session Data/IP-usage-Maori.txt <base directory>/Session Data/IP-usage-English.txt <base directory>/Session Data/IP-usage-Bilingual.txt <base directory>/Results/Zargument-exploratory-switching.txt <base directory>/Results/Zargument-usage-switching.txt <base directory>/Results/IP-exploratory-switching.txt <base directory>/Results/IP-usage-switching.txt
Additional Requirements:	The script takes the <base directory> as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.
Table A-3: Summary of split-sess-lang.pl.	

Name:	analyse-single-types.pl
Description:	Produces data about single request sessions.
Size of Script:	282 lines of code
Input Files:	<base directory>/Session Data/Zargument-singles.txt
Output Files:	<base directory>/Results/Single Request Data.txt
Additional Requirements:	The script takes the <base directory> as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.
Table A-4: Summary of analyse-single-types.pl.	

Name:	session-stats.pl
Description:	Produces large amounts of statistics about sessions. Produces files that contain apparent read times.
Size of Script:	1107 lines of code
Input Files:	<base directory>/Session Data/Zargument-exploratory.txt <base directory>/Session Data/Zargument-usage.txt <base directory>/Session Data/Zargument-exploratory-Maori.txt <base directory>/Session Data/Zargument-exploratory-English.txt <base directory>/Session Data/Zargument-usage-Maori.txt <base directory>/Session Data/Zargument-usage-English.txt <base directory>/Session Data/Zargument-usage-Bilingual.txt
Output Files:	<base directory>/Results/Exploratory Session Data.txt <base directory>/Results/Usage Session Data.txt <base directory>/Results/Exploratory Sessions in Maori Data.txt <base directory>/Results/Exploratory Sessions in English Data.txt <base directory>/Results/Usage Sessions in Maori Data.txt <base directory>/Results/Usage Sessions in English Data.txt <base directory>/Results/Bilingual Usage Sessions Data.txt <base directory>/Results/Reading Times/PageRead-ExploratoryMi.txt <base directory>/Results/Reading Times/PageRead-ExploratoryEn.txt <base directory>/Results/Reading Times/DocRead-UsageMi.txt <base directory>/Results/Reading Times/PageRead-UsageMi.txt <base directory>/Results/Reading Times/DocRead-UsageEn.txt <base directory>/Results/Reading Times/PageRead-UsageEn.txt <base directory>/Results/Reading Times/DocRead-UsageBi.txt <base directory>/Results/Reading Times/PageRead-UsageBi.txt
Additional Requirements:	The script takes the <base directory> as a command line argument. The script requires 'logs.pl' and 'analysis.pl'.

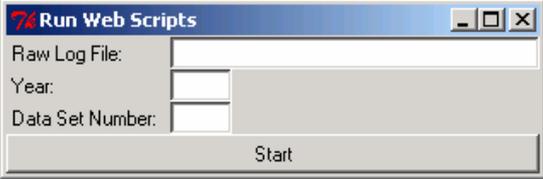
Table A-5: Summary of session-stats.pl.

Name:	logs.pl
Description:	A utility script containing functions for reading, manipulating and outputting raw cgi logs like those produced by the greenstone web interface. Equivalent functions are also available for the logs processed using the functions in this script.
Size of Script:	416 lines of code
Input Files:	
Output Files:	
Notes:	Required by all Perl scripts except 'renameFile.pl', 'renameLengths.pl' and 'gui.pl'.

Table A-6: Summary of logs.pl.

Name:	analysis.pl
Description:	A utility script for use in session analysis. Functions are available for reading in sessions from session logs, creating summary statistics, getting time and query term data.
Size of Script:	379 lines of code
Input Files:	
Output Files:	
Additional Requirements:	Requires logs.pl
Notes:	Required by all Perl scripts except 'logs.pl', 'process-raw-logs.pl', 'analyse-single-types.pl', 'renameFile.pl', 'renameLengths.pl' and 'gui.pl'.

Table A-7: Summary of analysis.pl.

Name:	gui.pl
Description:	<p>An interface to make using the set of scripts easier.</p> <p>Allows users to drag and drop the raw log, enter the year of the log, enter the data set number for this log then start to run all scripts.</p> 
Size of Script:	96 lines of code
Input Files:	
Output Files:	
Additional Requirements:	User must specify year of the script and what dataset it will have. All the other logs mentioned here must be in the same directory as this script.
<i>Table A-8: Summary of gui.pl.</i>	

Name:	renameFiles.pl
Description:	Simple script run to rename all the results files to a format specified by the file names.txt.
Size of Script:	76 lines of code
Input Files:	names.txt (in same directory as the log) (The files that are altered are in <base directory>/Results/) <base directory>/Results/General Data.txt <base directory>/Results/General Session Data.txt <base directory>/Results/Single Request Data.txt <base directory>/Results/Exploratory Session Data.txt <base directory>/Results/Usage Session Data.txt <base directory>/Results/Exploratory Sessions in Maori Data.txt <base directory>/Results/Exploratory Sessions in English Data.txt <base directory>/Results/Usage Sessions in Maori Data.txt <base directory>/Results/Usage Sessions in English Data.txt <base directory>/Results/Bilingual Usage Sessions Data.txt <base directory>/Results/Zargument-usage-switching.txt
Output Files:	(Assuming we pass a base number of 1) <base directory>/Results/ 1.1 – General Data.txt <base directory>/Results/ 1.2 – General Session Data.txt <base directory>/Results/ 1.3 – Single Request Data.txt <base directory>/Results/ 1.4 – Exploratory Session Data.txt <base directory>/Results/ 1.5 – Usage Session Data.txt <base directory>/Results/ 1.6 – Exploratory Session in Maori Data.txt <base directory>/Results/ 1.7 – Exploratory Session in English Data.txt <base directory>/Results/ 1.9 – Usage Sessions in Maori Data.txt <base directory>/Results/ 1.10 – Usage Sessions in English Data.txt <base directory>/Results/ 1.11 – Bilingual Usage Sessions Data.txt <base directory>/Results/ 1.12 – Switching Data.txt
Additional Requirements:	The base number must be passed to the script as a command line argument.
Notes:	files are renamed as follows: <base number>.<secondary number> - <description>.txt Where <base number> is a number passed to this script and <secondary number> and <description> are specified in names.txt
Table A-9: Summary of renameFiles.pl.	

Name:	renameLengths.pl
Description:	Simple script which is run to rename all the length files to a format specified in the script.
Size of Script:	56 lines of code
Input Files:	(All the files to be renamed are in /Results/Reading Times/) <base directory>/Results/Reading Times/PageRead-ExploratoryMi.txt <base directory>/Results/Reading Times/PageRead-ExploratoryEn.txt <base directory>/Results/Reading Times/DocRead-UsageMi.txt <base directory>/Results/Reading Times/PageRead-UsageMi.txt <base directory>/Results/Reading Times/DocRead-UsageEn.txt <base directory>/Results/Reading Times/PageRead-UsageEn.txt <base directory>/Results/Reading Times/DocRead-UsageBi.txt <base directory>/Results/Reading Times/PageRead-UsageBi.txt
Output Files:	(Assuming we have a base number of 1) <base directory>/Results/Reading Times/1.6 - PageRead - ExploratoryMi.txt <base directory>/Results/Reading Times/1.7 - PageRead - ExploratoryEn.txt <base directory>/Results/Reading Times/1.9 - DocRead - UsageMi.txt <base directory>/Results/Reading Times/1.9 - PageRead - UsageMi.txt <base directory>/Results/Reading Times/1.10 - DocRead - UsageEn.txt <base directory>/Results/Reading Times/1.10 - PageRead - UsageEn.txt <base directory>/Results/Reading Times/1.11 - DocRead - UsageBi.txt <base directory>/Results/Reading Times/1.11 - PageRead - UsageBi.txt
Additional Requirements:	The base number must be passed to the script as a command line argument.
Notes:	files are renamed as follows: <base number>.<secondary number> - <doc OR page> - <Session><Lang>.txt Where <base number> is a number passed to this script, <secondary number> is specified in this script and should relate to the session stats file numbers—that is the usage English stats file should have the same numbers as the doc and page read files relating to it. <doc OR page> is either DocRead or PageRead depending on whether this is a page read length or document read length file. <Session> is either Usage or Exploratory depending on what the type of session file is. <Lang> is either En, Mi, or Bi depending on whether this session log file was an English (En), reo Māori (Mi) or Bilingual (Bi) session.
Table A-10: Summary of renameLengths.pl.	

The Perl scripts *analyse-single-types.pl* and *session-stats.pl* generated the majority of the data used in the Transaction Log Analysis. Examples of the data and statistics generated include:

i) General Data

- Counts of requests in reo Māori and in English.
- Counts of unique users in reo Māori and in English.
- Counts of total requests and total unique users.
- Counts of pages requested; home page, about page, explanation page, help page, preferences page, gsdl page and unknown pages.
- Counts of documents requested from search, series, date and unknown.
- Counts of document types requested, text, full image, preview and other.

ii) Session Length Summary Data

- Statistical data on length of session in time includes; mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on length by requests includes; mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.

iii) Session Length Listing

- Listing of session counts by length of sessions in minutes.
- Listing of session counts by length of sessions in requests.

iv) Chronological Data

- Counts of requests per month.
- Counts of requests per day of week.
- Counts of requests per hour of the day.
- Counts of sessions per month.
- Counts of sessions per day of week.
- Counts of sessions per hour of the day.

v) Requests per Client

- Listing of Z arguments by count of sessions and counts of requests.
- Listing of IP addresses by count of cookies.
- Listing of IP addresses by count of requests.

vi) Session Query Data

- Counts of queries and query pages.

- Statistical data on queries per session includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Counts of queries that used search operators, e.g. Boolean, quotation etc.
- Counts of queries that changed the query arguments, including maximum query requests, search type (Some/All) and case matching.
- Counts of sessions that had no queries, 1 query, 2 queries, 3 queries, 4 queries, 5 queries and 6 or more queries.
- Statistical data on number of terms per query includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Count of queries with 1 query term, 2 query terms and 3 or more query terms
- Statistical data on number of documents viewed after a query includes; mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on number of text documents, full size image documents and preview images viewed after a query.
- Statistical data on time (in seconds) between document views after a query includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on time (in seconds) between text documents viewed after a query, between full size images viewed after a query and between preview images viewed after a query.

vii) Listing of Query Terms

- Listing of query terms by number of times submitted and number of unique users submitting.

viii) Documents Viewed

- Listing of documents viewed by count of requests and count of unique user.

ix) Times Between Requests and Views

- Statistical data on time between successive page requests includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.
- Statistical data on time between successive document requests includes mean, mode, standard deviation, lowest, lower quartile, median, upper quartile, highest and count.

Appendix B Hawaiian Query Terms

The query terms submitted to the Hawaiian Nūpepa search engine were collated but not listed, in Chapter 6 Hawaiian Language Usage of the Niupepa Digital Library. The query terms were not listed because it was felt that the analysis of these terms was beyond the scope of the thesis. However, the query terms are listed in this appendix for the benefit of those who would like to analyse them.

In total there were 58,846 query terms submitted to the Hawaiian Nūpepa digital library in 2005. There were 18,478 query terms used in 11,232 queries undertaken in usgHaw sessions, 22,026 query terms used in 14,966 queries undertaken in usgEn sessions, and 14,342 query terms used in 9,428 queries undertaken in usgBi sessions.

Two methods can be used to determine the most popular query terms. The first method is simply a count of each time a term, or string of terms, was submitted to the search engine. The second method is a count of unique users who submitted the search term or search string. Terms are displayed sorted by unique user as this reduces the effect of users who are making repeat searches on the one term and gives a better indication of the variety of topics for which clients were searching.

The most common query terms submitted are displayed in *Table B-1*. The number of times that the term was requested is in square brackets, the number of unique users requesting that term is in curved brackets.

Most Common Hawaiian Nūpepa Query Terms in 2005

	usgHaw	usgEn	usgBi
1	hula [30] (17)	Pahau [58] (24)	liliuokalani [23] (11)
2	bush [23] (16)	kalawaia [39] (20)	hula [31] (9)
3	aloha [24] (15)	kamehameha [24] (18)	kamehameha [13] (7)
4	liliuokalani [24] (13)	kahoalii [55] (17)	kealoha [7] (7)
5	kanikau [33] (12)	bush [34] (15)	Liliuokalani [11] (5)
6	welina [30] (10)	naholowaa [25] (15)	annexation [8] (5)
7	kohola [28] (10)	Kamehameha [25] (14)	kalakaua [7] (5)
8	mokaulele [18] (10)	hula [26] (12)	silva [6] (5)
9	Kaupo [36] (9)	kaaawa [14] (12)	kalawaia [47] (4)
10	kaiulani [20] (9)	manoa [13] (11)	Wond [34] (4)
11	Pele [17] (9)	liliuokalani [12] (11)	Makuaole [30] (4)
12	kumulipo [14] (9)	kaneohe [12] (10)	kaumualii [18] (4)
13	maui [13] (9)	holua [24] (9)	welina [16] (4)
14	kaleponi [35] (8)	nahinu [19] (9)	napulou [15] (4)
15	kealohi [25] (8)	napulou [23] (8)	Queen Emma [14] (4)
16	kaao [18] (8)	Kolomoku [23] (8)	Kamehameha [14] (4)
17	palapala sila nui [16] (8)	hana [14] (8)	mary kealoha [9] (4)
18	kalo [14] (8)	mahele [9] (8)	kamapuaa [9] (4)
19	kaluaikoolau [14] (8)	kawaihau [22] (7)	kauhane [9] (4)
20	Kamapuaa [14] (8)	Purdy [19] (7)	heenalu [8] (4)
21	mookuauhau [13] (8)	kamakau [18] (7)	molokai [7] (4)
22	mano [9] (8)	naki [15] (7)	mookuauhau [7] (4)
23	Keauhou [25] (7)	heenalu [13] (7)	hana [6] (4)
24	Kahekili [23] (7)	kekuhaupio [12] (7)	kaaawa [6] (4)
25	Kawelo [19] (7)	hawaiian language [12] (7)	bush [5] (4)
26	kawelo [12] (7)	kahele [12] (7)	
27	Liliuokalani [11] (7)	kanikau [10] (7)	
28	koolau [9] (7)	overthrow [9] (7)	

Table B-1: Query terms with [number of requests] and (number of unique users) submitted by usgHaw, usgEn and usgBi clients to the Hawaiian Nūpepa digital library.

Appendix C Interview Forms

The interview procedure is described in §7.3. The interviews began with an explanation of the purpose of the research. The participants were then shown, and had explained to them, a Bill of Rights and a Research Consent Form which they were asked to sign if they were happy to be interviewed. This consent form was also signed by the interviewer/researcher.

The Bill of Rights form, shown in *Figure C-1*, consists of two parts. The first part, written only in te reo Māori, greets the participants in a traditional manner and expresses gratitude for taking time to be involved with this research. The second part, written only in English, lists nine rights that each participant was entitled to throughout all stages of the interview procedure.

The Research Consent Form, shown in *Figure C-2*, gives some background information about the research and the interview procedure. It then details what the participant has agreed to and provides a place for the signatures of the participant and the researcher. Two copies of the form were signed; one was retained by the participant, the other by the researcher.

Kaupapa Rangahau: te Kohikohinga Niupepa/Nūpepa

He hōnore, he korōria ki runga.
He maunga rongō ki te mata o te whenua.
He whakaaro pai tētehi ki tētehi.

Rire, rire, haul Pai mārire.

E ngā mana e ngā reo e ngā tapu e noho atu nā, tēnei te tuohu te tuku mihi tēnā koutou katoa.

E tuku mihi ana hoki ki a koe kua whai wā koe ki te noho mai ki te āwhina i te kaupapa, ki te whakamahi i te Kohikohinga Niupepa, ki te kōrero hoki mō ētehi āhuatanga kua kitea e koe. Hei aha? Hei arahi hei āwhina i ngā kōkiri whakamua o tō tātou reo rangatira i roto i ngā kaupapa hāngarau pēnei nā. Nō reira ka nui rawa te mihi, tēnā hoki koe.

Kāti. Ka kite atu e rua ngā kaupapa kei runga i ngā whārangi nei. Tuatahi he whakamārama i ōu mōtika, arā ngā *Participant's Rights*, tuarua he puka whakaaetanga māu mō te mahi rangahau nei, arā te *Research Consent Form*.

Tēnā koa, pānuitia ngā kōrero nei, ā, mehemea e pai ana hainatia te puka whakaaetanga.

Mehemea, he pātai, he āwangawanga, kōrero mai koa.

Nā tō pononga noa nei,

nā Te Taka Keegan
Pūkenga, Kairangahau, Tāri Rorohiko
Te Whare Wānanga o Waikato

Research Participant's Bill of Rights

The following is a list of your rights if you participate in this usage evaluation of Niupepa/Nūpepa.

As a research participant, you have the right:

1. To be treated with dignity and respect at all times.
2. To choose not to be involved in this evaluation.
3. To withdraw from this evaluation at any time.
4. To know that all information you provide in this evaluation is anonymous, and will be kept strictly confidential.
5. To be fully and clearly informed of all aspects of the research prior to becoming involved in it.
6. To enter into clear, informed, and written agreement with the researcher prior to becoming involved in this activity. You should sense NO pressure, explicit or otherwise, to sign this contract.
7. To be treated with honesty, integrity, openness, and straightforwardness in all phases of the website usage evaluation.
8. To expect that your personal welfare is protected and promoted in all phases of the research, including knowing that no harm will come to you.
9. To access any information that has been collected about your evaluation and to remove this information if you are no longer comfortable with it being used in this research.

Figure C-1: Bill of Rights for Research Participants.

Research Consent Form

** Please take the time to read and understand this form.*

Research Project Title: Niupepa/Nūpepa Website Usage Evaluation

Researcher: Te Taka Keegan

Evaluation Purpose: The purpose of this evaluation is to investigate bilingual usage of the Niupepa and Nūpepa Websites. Web site log analysis has highlighted different usage characteristics of people who access the site in Māori/Hawaiian compared with people who access the site in English. The intention of this evaluation is to try to determine why some of those differences are occurring.

Participant Recruitment and Selection: Potential users of the Niupepa/Nūpepa website are solicited by the researcher and colleagues and invited to participate in this evaluation.

Procedure: This evaluation consists of three distinct stages. First you will be introduced to the purpose and procedure of the Niupepa/Nūpepa website evaluation. Next you will be asked to use the Niupepa/Nūpepa website to investigate some topic of interest to you. Finally you will be asked some questions about your experience with using the Niupepa/Nūpepa website.

Explaining the evaluation and allowing you to read the information is expected to take about 15 minutes, using the website is expected to take about 20 minutes, and the final questioning stage is expected to take about 20 minutes. Consequently the evaluation should take approximately 1 hour.

It is your choice as to whether this evaluation is conducted in te reo Māori or in English.

Data Collection: Your usage of the website and the questions you are asked afterwards will be videoed and recorded to DVD. This is to provide an accurate record of your activity and responses. This DVD will not be publicly available. It will not be privately available outside the purposes of this research without prior written consent from the participants. The DVD will be kept secure in my office, in G.2.07 at the University of Waikato. On completion of the research the DVDs will be destroyed.

Confidentiality: Confidentiality and participant anonymity will be strictly maintained. You will not be identified in any publication or dissemination of research findings.

Likelihood of Discomfort: There is no likelihood of discomfort or risk associated with participation.

Finding out about Results: You can find out about the results of the study by contacting the researcher, Te Taka Keegan by email (tetaka@waikato.ac.nz) after December 2006.

Agreement

Your signature on this form indicates that you have understood to your satisfaction the information regarding participation in this research and agree to participate as a participant. In no way does this waive you legal rights nor release the researchers from their legal and professional responsibilities. You are free to withdraw from the study at any time. Your continued participation should be as informed as your initial consent, so you should feel free to ask for clarification or new information throughout your participation. If you have further questions concerning matters related to this research, please ask the researcher.

Participant

Date

Investigator/Witness

Date

**You may keep a copy of this consent form for your records for future reference should you need it.*

Figure C-2: Research Consent Form.

