INDIGENOUS LANGUAGE USAGE IN A BILINGUAL INTERFACE – TRANSACTION LOG ANALYSIS OF THE NIUPEPA WEBSITE

Te Taka Keegan,
Computer Science Department
University of Waikato
Private Bag 3105,
Hamilton, New Zealand
64 7 838 4420
64 7 858 5095 (fax)
tetaka@cs.waikato.ac.nz

Dr. Sally Jo Cunningham,
Computer Science Department
University of Waikato
Private Bag 3105,
Hamilton, New Zealand
64 7 838 4402
64 7 858 5095 (fax)
sallyjo@cs.waikato.ac.nz

Prof. Mark Apperley,
School of Computing & Mathematical Sciences
University of Waikato
Private Bag 3105,
Hamilton, New Zealand
64 7 838 4528
64 7 858 5095 (fax)
m.apperley@waikato.ac.nz

INDIGENOUS LANGUAGE USAGE IN A BILINGUAL INTERFACE – TRANSACTION LOG ANALYSIS OF THE NIUPEPA WEBSITE

ABSTRACT

In this article we investigate the extent and characteristics of use of the Māori language, the indigenous language of Aotearoa (New Zealand), in a large bilingual website. We used transaction log analysis to investigate whether Māori was utilised by users of the website and how usage characteristics differed between users of Māori and users of the more commonly spoken English language. We found that Māori language was used in one quarter of all active sessions, and that in these sessions users were more likely to browse the website, whereas users working in the non-indigenous English were more likely to use the search facility. We also identified a new category of user of bilingual websites: the bilingual user.

INTRODUCTION

To analyse indigenous language participation in a website one must be aware of how many potential users there are of the website in the indigenous language. Approximately 14% of the total resident population of Aotearoa (New Zealand) is Māori and about 1 in 4 of these is able to converse in te reo Māori (the Māori language) (Te Puni Kōkiri 2003). When we consider that 65% of Māori have never accessed the Internet, the potential users of a te reo Māori language interface is perhaps 1-2% (40,000-80,000) of the population of Aotearoa. In contrast potential English-speaking users of the website in Aotearoa represent approximately 51% of the population (2,040,000) (Te Puni Kōkiri 2001).

However, as well as potential user numbers, one must also consider the content of the website and whether the information available favours a particular language; in this case the odds shift back to the indigenous language. The Niupepa (Māori for newspaper) website makes available over 17,000 historic newspaper pages collected from 35 periodicals published between the years 1842 and 1933, the time period when most te reo Māori periodicals were published. The number of publications in te reo Māori diminished after this time, and it was not until the latter part of the 20th century when a resurgence of interest in Māori language led to a renewed interest in texts written in Māori. In 1996 the Alexander Turnbull Library published 'Niupepa 1842-1933', a 407 page microfiche set containing facsimiles of the pages of these periodicals that had been preserved in libraries throughout Aotearoa. In 2002 the Computer Science Department at the University of Waikato made the collection available on the Internet through he bilingual Niupepa website (see Apperley et al 2002), on which the study in this article is based. As 70% of the pages are exclusively in te reo Māori, and 27% are in both English and Māori, the content is most suited to users who are literate in te reo Māori.

While there have been articles written describing minority language use on bilingual websites (see Cuncliff 2003, IBIS 2000 and Warshchauer et al 2002) and publications produced on transaction log analysis (see Jones et al 2000, Koch et al 2005, and Pikow 1997) this article is unique in that transaction log analysis is used to determine indigenous language usage and characteristics in a bilingual website.

Limitations of this type of analysis

While every endeavour has been made to ensure that the data used is as accurate and as meaningful as possible, log file analysis does have shortcomings, in particular the effect of

Web caches. A Web cache intervenes between a Web server and a Web client; it will notes requests from the client and saves copies of the server's responses. If it detects a request to which it already has a copy of the response, it will supply the response directly and the request is not passed on to the original server. This saves time and reduces network traffic. However, as the original server (in this case the Niupepa server) does not receive the request, it is not recorded in its Web log file and so this user action is not included in the transaction log analysis.

There are two main types of Web cache; a browser cache, which is handled by a user's browser software, and a network (or proxy) cache, which is configured within a local area network. Both types of caches have the effect of masking repeated requests for the same data from a single user appearing in the transaction log; however, a network cache will also mask duplicated requests from different users within the same local-area network.

Other limitations of log file analyses include false hits due to Web robot activity, false hits caused by server upgrades and maintenance, and the inability to accurately delimit individual user sessions. We describe our efforts to deal with these limitations below. The analysis presented in this article, as with all transaction log analyses, must be viewed with the knowledge that not all user activity will appear in the Web logs because of the effect of Web caching.

The Niupepa Collection

The Niupepa Collection is currently delivered by the Greenstone software of the New Zealand Digital Library (NZDL) at: www.nzdl.org/niupepa (for a description of the Greenstone software see Witten and Bainbridge 2002). For a comprehensive explanation regarding the process of delivering the Niupepa on the Web, see Apperley et al (2002). The software

provides three methods of accessing the newspapers in the collection: full text search, browsing the newspapers by series and issues, and browsing the newspapers by a timeline.

Picture 1: The Niupepa Website with the interface set to teo reo Māori

The newspaper pages themselves may be viewed as extracted text, or in either of two facsimile forms: a low resolution image that downloads quickly for previewing, or a high resolution image that takes longer to download but which it is possible to read on the screen. While the content is a mixture of Māori and English text (primarily Māori; see earlier), the interface to the collection can be presented in either the Māori or English, and this language can be switched at any time. The default (starting) language of the collection is set to Māori.

GATHERING THE DATA

All user activity on the Niupepa website is logged. Every individual access event or 'hit' is recorded with details such as the page requested, the language currently set in the interface,

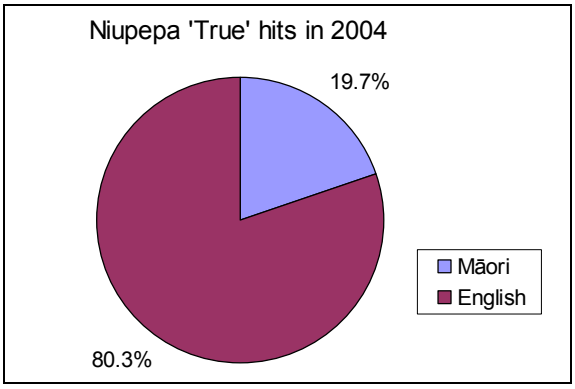
the time of the request, the nature of the request, the previous action, the IP address of the requester and other various user preferences that are currently set.

The data used for analysis in this report is from a log of 187,215 hits on the Niupepa site for the year 2004, recorded from 12:17am January 1st 2004 to 11:41pm December 31st. This average of more than 500 hits per day shows the strong usage of the site. The log represents usage from within New Zealand only, as all off-shore requests were handled by a Web server located at the University of Lethbridge in Alberta, Canada, and so are not included in the log.

The raw Niupepa transaction log data was filtered to remove unwanted hits. These included known Web robot hits (338), hits where the IP address was not defined (495), hits from the local research team (1565), and hits where the interface language setting could not be determined (3578).

Filtered Hits by language

The resulting filtered Niupepa log totalled 181,239 hits, comprising of 145,596 hits (80.3%) where the interface language was set to English, and 35,643 hits (19.7%) where the interface language was set to Māori, as shown in Graph 1.



Graph 1: Total Niupepa Filtered hits in 2004 by language

Defining Sessions

To improve the usefulness of the data and to make it more relevant to the actual usage of the Niupepa website, hits in the filtered log were grouped into sessions, where we define a session as a sequence of two or more hits originating from a single user, with no more than a 60 minute gap between successive hits. Log analyses usually define sessions as having successive hits separated by no more than 30 minutes (see Pikow 1997, p1348); however, because Niupepa users might be expected to spend extended periods examining a single newspaper page, and because the gap between hits might be exaggerated by Web caching (which could mask intermediate hits) we felt that for our analysis it was appropriate to extend the maximum spacing to 60 minutes.

In order to identify sessions, we needed to associate hits with individual users. For this purpose, two methods were available. The first is to identify users by the IP address of the computer responsible for the request, as recorded in the transaction log. The second method involves the use of cookies (see Pitkow 1997, p1343); when a user begins a session a cookie is created on their machine which includes the IP address of the computer and the time that the cookie was created. This information is included as the z argument in each transaction log record. Note that the z argument cannot distinguish between two people using the same login to a given computer (for example, if two or more people use a public library computer to access a collection).

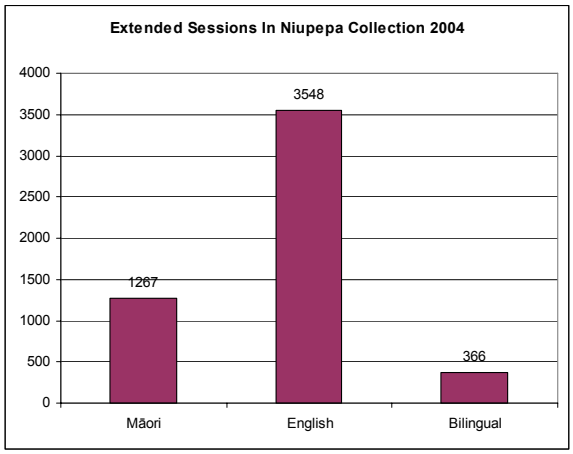
Both techniques have their drawbacks. The IP address is machine specific, but not user specific; consequently if multiple users share one computer in sequence then several quite distinct sessions with different users might be seen as a single session in the log. People connecting through a shared dialup connection may also have a common IP address, and

again multiple users' activities might be logged as a single session. Using the z argument to define users greatly reduces the problem of multiple users being recorded as a part of the same session; however session activity of users who disable cookies will not be logged. We chose to use the z argument to identify sessions, as although this reduced the data set by 17% (sessions where cookies were disabled), we were more confident that each identified session related to an individual user.

The sessions were identified and then classified into one of two groups. *Exploratory* sessions were defined as those where the user accessed only the home page, the help page, and/or the preferences page. No documents in the collection were accessed and no searches were undertaken. *Extended* sessions were defined as those involving queries and/or retrieval of documents from the Niupepa collection. Our objective in analyzing sessions was to determine if there were observable differences between those carried out in te reo Māori and those carried out in English. However closer examination of the session data revealed that there was a third category of user: the bilingual users, who conducted their sessions using a mix of the Māori and English interfaces.

For the purposes of our analysis, we defined a te reo Māori session as one that used the Māori interface for at least 80% of its logged events, and did not involve more than two interface language switches. An English language session was defined as a session that used the English interface for at least 80% of its logged events and did not involve more than two interface language switches. We defined a bilingual session as one that involved three or more interface language switches and/or where there was at least 20% of the log activity in each language.

Of the 5653 individual sessions identified, 472 were exploratory. Because this represents such a small number of hits there is little to be gained from the analysis of exploratory sessions. There were 5181 extended sessions: 1267 te reo Māori sessions, 3548 English language sessions and 366 bilingual sessions (See Graph 2). The next section discusses the analysis of the extended sessions.



Graph 2: Classification of Extended Sessions

ANALYSIS OF THE EXTENDED SESSION

Table 1 shows the general statistics for the extended sessions, grouped according to the three language categories previously defined. Most of the sessions, 3548 (68.5 %), are in English, with 1267 (25.4%) in Māori and the remaining 366 (7.1%) making significant use of both languages. In general terms, the English sessions are more sustained, with longer sessions and more hits per session.

	Māori	English	Bilingual
sessions:	1267	3548	366
total session %:	24.5%	68.5%	7.1%
unique users:	2174	1057	1239

page hits:	29055	108479	7845
mean (hits):	22.9	30.6	21.4
median (hits):	7	15	9
longest session (hits):	408	679	182
shortest session (hits):	2	2	2
std deviation (hits):	33.0	44.7	28.0
mean (min):	18.7	25.2	16.3
median (min):	7	8	5
longest session (min):	390	711	241
shortest session (min):	<1	<1	<1
std deviation (min):	31.8	45.4	27.6

Table 1: General Statistics for Extended Sessions

Extended Session Analysis – Accessing and Viewing of Newspaper Pages

The transaction log records indicate whether the newspaper pages accessed were the result of a search, or by browsing by newspaper series or date. This analysis is shown in Table 2, where it can be seen that in all three categories, more document pages were retrieved by searching than by browsing.

	Māori		**English**		**Bilingual**	
pages viewed from search:	14564	72.4%	64125	83.3%	2686	67.8%
pages viewed from series:	3782	18.8%	6838	8.9%	761	19.2%
pages viewed from date:	1532	7.6%	4818	6.3%	363	9.2%
other:	234	1.2%	1194	1.6%	149	3.8%

Table 2: Niupepa pages retrieved by Search, by Series and by Date

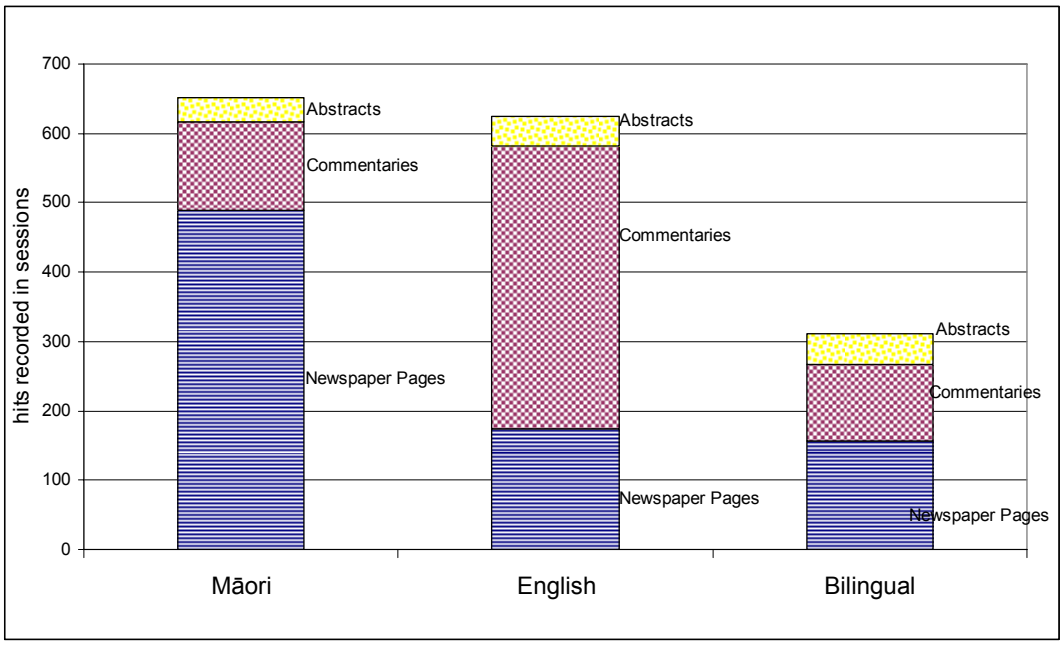
Table 3 shows the manner in which these retrieved pages were viewed; in the form of extracted text, as a preview facsimile page, or as a full-size facsimile page. Viewing the extracted text is the most common way to view a page, but then this is the default display

following a search. However the users in the Māori session show a greater tendency to view the full facsimile image (21.5%) than those in the English sessions (5.6%) or the bilingual sessions (9.9%).

	Māori		English		Bilingual	
text pages viewed:	14724	73.2%	50746	65.9%	2740	69.2%
full images viewed:	4310	21.4%	4300	5.6%	391	9.9%
preview images viewed:	1014	5.0%	21752	28.3%	801	20.2%
undefined:	64	0.3%	177	0.2%	27	0.7%

Table 3: Viewing of Text Pages, Full Images and Preview Images

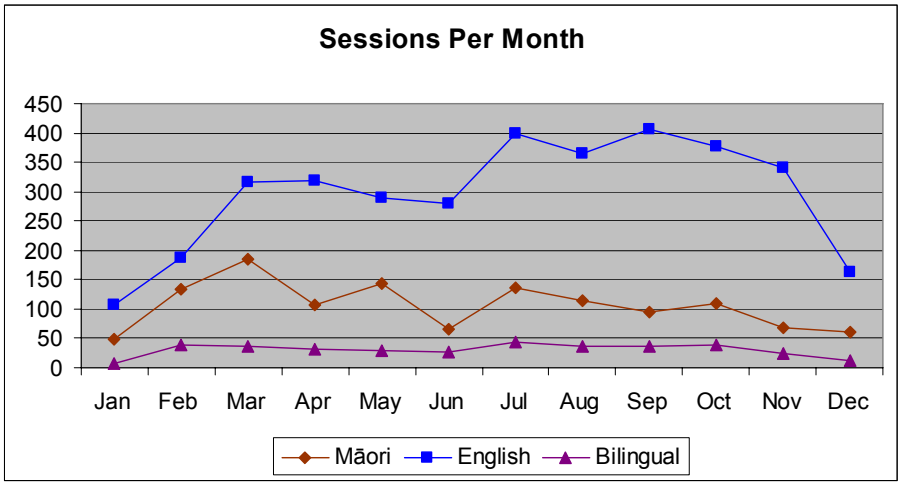
As well as delivering newspaper pages, the Niupepa Collection makes available two other types of information: commentaries which include bibliographic details, background, subject matter and accessibility, written mostly in English; and abstracts written in English which summarise the periodicals. If we look at the 15 most popular pages by hit count (Graph 3) in each language category, it can be clearly seen that Māori session users most commonly access actual newspaper pages, while English session users more commonly access the commentary information. This seems logical, as the users are accessing information that is in the chosen language of the session.



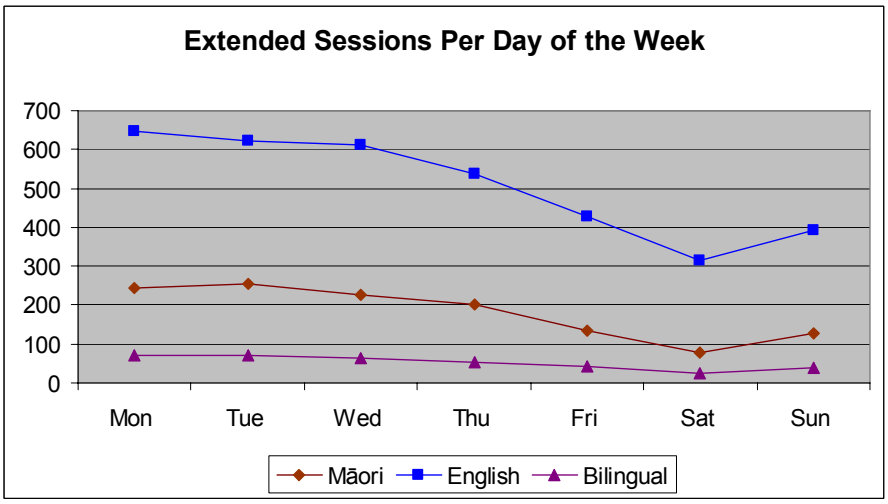
Graph 3: Top 15 Page Type by Hit Count

Session Analysis – Time Statistics

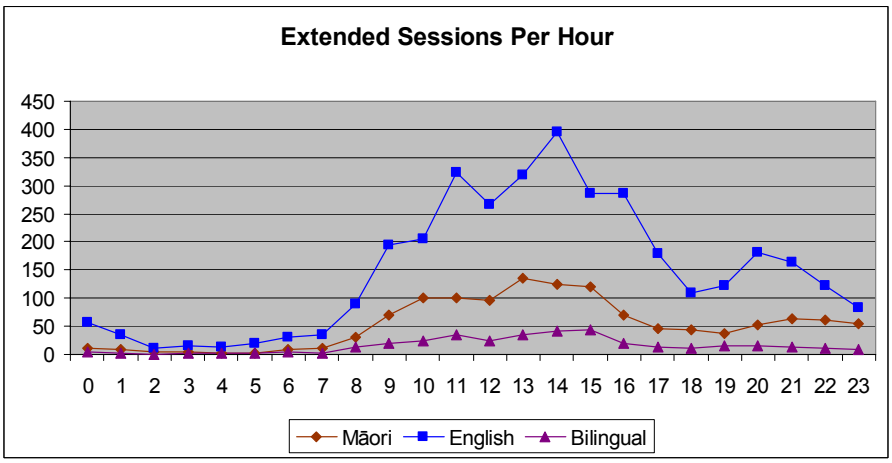
The extended sessions were also plotted against time (Graph 4). This shows lower number of sessions in the New Zealand summer holiday months of December and January. Breaking this down to daily usage (Graph 5) we see that while the sessions are reasonably consistent in the start and middle of the week they taper off towards the end of the week. There are more sessions occurring on a Sunday than on a Saturday. An examination of hourly exploratory sessions (Graph 6) shows that the majority of the sessions are occurring during the 9-5 work hours, with some further activity occurring in the evenings.



Graph 4: Extended Sessions per Month 2004



Graph 5: Extended Sessions per Day 2004



Graph 6: Extended Sessions per Hour 2004

Extended Session Analysis – Searching

Statistics on searching conducted in extended sessions are presented in Table 4 This shows an average of three or four searches per session, and that 30-40% of all sessions involves three or more searches.

	Māori	English	Bilingual
number of searches:	4289	16888	1127
average searches per session:	3.4	4.8	3.1
Std Deviation of Queries:	3.06	2.45	2.66
0 search sessions:	35.4%	24.8%	37.2%
1-2 search sessions:	29.4%	30.7%	31.7%
3+ search sessions:	35.2%	44.4%	31.1%

Table 4: Extended Session Searching Activity

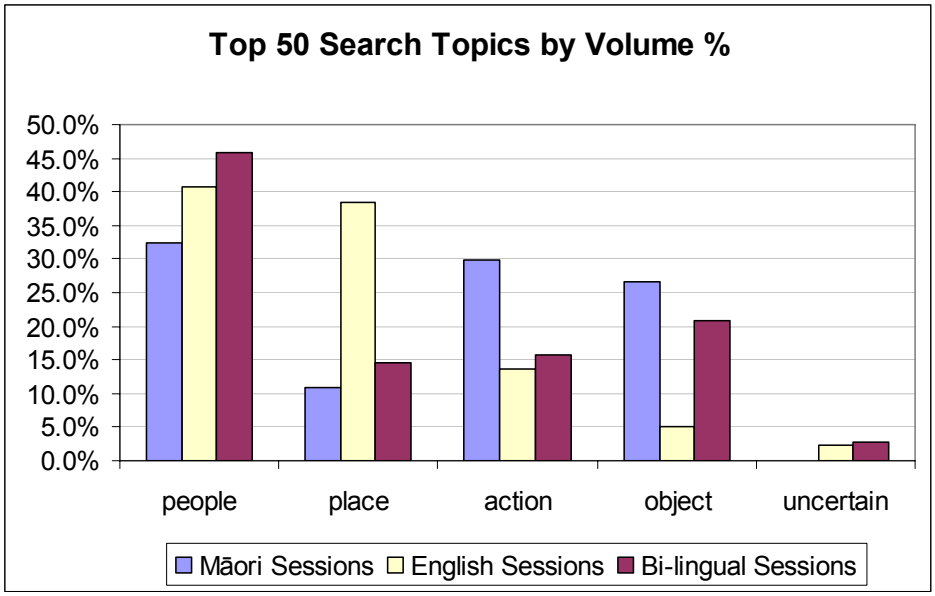
The number of terms used in searches is similar across the three extended session types averaging just under two terms per search for all 3 session types (Table 5). Approximately half of all searches undertaken use just the one search term.

	Māori	English	Bilingual
average terms per search:	1.8	1.9	1.9
STD DEV of search terms:	0.33	0.36	0.45
searches with 1 term:	53.3%	47.0%	51.9%
searches with 2 terms:	26.2%	32.6%	27.4%
searches with 3+ term:	20.5%	20.4%	20.7%

Table 5: Extended Session Searching Terms

We examined the top 50 searches that were submitted in the extended sessions using the different language interfaces. The search terms were manually examined and sorted into categories; people e.g Apanui, place e.g. Parihaka, action e.g. poroporoaki (farewell), object

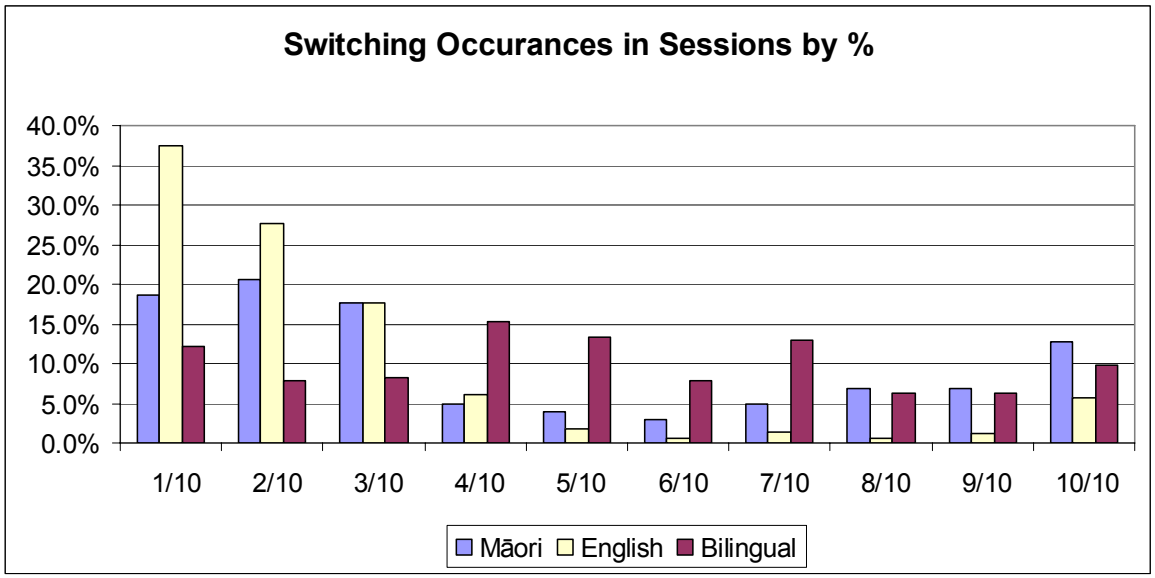
e.g. Matariki (star constellation), and uncertain e.g. Ngarakau (could be a name of a person or a place). The Māori extended session searches were reasonably spread across the four topics, while the English extended sessions were primarily concerned with people and places (Graph 7). The bilingual extended sessions showed results that were between the Māori and English sessions.



Graph 7: Top 50 Search Topics

Extended Session Analysis – Language Switching

We also carried out an analysis to find out at what point in the sessions users chose to switch the interface language. The results are displayed in Graph 8, and clearly show that language switching in the English sessions occurs mostly (82.8%) in the first 30% of the session. In te reo Māori sessions language switches is more evenly distributed with only 28.3% in the first 30% of the sessions. The language switching in the bilingual sessions is distributed evenly with no clear peak or trough. This gives us an indication that this particular user group will switch the language of the interface at any point in the session.



Graph 8: Position where language switching occurs in a Session by Percentage

CONCLUSIONS

The transaction log analysis of the bilingual Niupepa website has uncovered three important findings that have relevance to the usage of information technologies by indigenous people.

The first and most important conclusion drawn is that even though the potential number of users of a bilingual website in an indigenous language such as Māori is very low (1-2% of Aotearoa's population) the website is still used considerably in the indigenous language (approximately 1 in 4 of all active sessions). The decision to offer the Niupepa website in Māori has been justified by active use of the website in te reo Māori. When we consider that 99% of all Māori speakers are in fact Māori then the usage statistics inform us that this particular website has active participation by Māori people.

The second finding is that session analysis suggests some important usage differences between indigenous and non-indigenous users. The Māori session users show a greater

tendency to access the information by browsing, to download full size images (presumably for online reading), and when undertaking searching tend to search for a wide range of topics (as opposed to English searching which is predominantly for names and places). While some of these usage differences may be explained by the language of the content, it does appear that indigenous users of this bilingual website are using a wider variety of information retrieval and display options than non-indigenous users.

The third significant outcome of this research is the discovery of another category of user that has not previously been considered where analyzing usage of a bilingual website. This is the bilingual user, a user who conducts a significant proportion of the session in each language. This user is likely to switch languages at any time during a session. A characteristic noted of this user is a higher use of browsing strategies, taking advantage of browser material available in both languages, and a subsequent lower dependence on the usage of the search engine.

REFERENCES

Apperley M. D., Keegan T. T., Cunningham S. J., & Witten, I. H. (2002). Delivering The Māori Newspapers on the Internet in Curnow J, Hopa N, McRae J (ed.s), Rere Atu Taku Manu! Discovering History Language and Politics In The Māori Language Newspapers. Auckland University Press. Pages 211-36.

Cuncliffe, D. (2003) Promoting minority language use on bilingual Web sites. In: proceedings of the 1st Mercator International Symposium, Aberystwyth, April 2003. Available on line at: http://www.aber.ac.uk/~merwww/english/events/mercSym_03-04-08.htm

IBIS (2000). Interfaces for Bilingual Information Systems project reports. Available on line at: http://weblife.bangor.ac.uk/ibis/

Jones S., Cunningham S. J., McNab R. J. & Boddie S. (2000). A transaction log analysis of a digital library. In International Journal on Digital Libraries 3(2) 152-169.

Koch T., ArdöA., Golub, K. (2004). Log Analysis of User Behaviour in the Renardus Web Service. Poster presentation at Libraries in the Digital Age Conference, Dubrovnik and Mljet, Croatia.

James Pitkow, (1997) In search of reliable usage data on the WWW, Selected papers from the sixth international conference on World Wide Web, p.1343-1355, Santa Clara, United States

Te Puni Kōkiri (2001). Māori Access to Information Technology. Te Puni Kōkiri, Wellington, New Zealand.

Te Puni Kōkiri (2003). Speakers of Māori within the Māori Population. Te Puni Kōkiri, Wellington, New Zealand.

Warschauer, M., El Said G. R., & Zohry, A. (2002). Language Choice Online: Globalization and Identity in Egypt. In Journal of Computer-Mediated Communication 7(4)

Witten, I. H., & Bainbridge, D. (2002). How to Build a Digital Library. Morgan Kaufmann. San Francisco, CA.

Terms and Definitions

Active Sessions: A session is defined as the series of requests that a Web user asks from a website in a given time period (in this case 1 hour). An active session is a session where the user does more than just look at the home and explanatory pages but either undertakes some searches or browses some of the content of the website.

Cookies: A cookie is a file on a Web user's hard drive that is used by Web sites to record data about the user.

Facsimile: A photographic reproduction of a page that is as true to the original as possible.

IP address: A 32-bit number that identifies each sender or receiver of information that is sent across the Internet. An IP address has two parts: the identifier of a particular network on the Internet and an identifier of the particular device (which can be a server or a workstation) within that network.

Transaction Log Analysis: An analysis that is undertaken on the data that has been recorded in a Web log. A Web log will record all requests for pages, often called hits, that are made to a Web server.

Microfiche: A small sheet (4" x 6") containing microfilmed images of pages, read with a microfilm reader. Many pages of text fit onto a single fiche, and their major advantage is in saving shelf space.

Web client: A Web browser is a software package that enables a user to display and interact with documents hosted by Web servers.

Web server: A Web server is a program that serves the files that form Web pages to Web users. Every computer on the Internet that contains a Web site must have a Web server program.

Web cache: A Web cache fills requests from the Web server, stores the requested information locally, and sends the information to the client. The next time the Web cache gets a request for the same information, it simply returns the locally cached data instead of searching over the Internet, thus, reducing Internet traffic and response time.

Web robot: Also known as a Web Wanderer or Web Spider, it is a program that traverses the Internet automatically by retrieving a document, and recursively retrieving all documents that are referenced. The data recorded in transaction logs by this type of activity is termed 'non-human' and can lead to misleading results so consequently must be removed when undertaking transaction log analysis.