LANGUAGE SWITCHING IN A DIGITAL LIBRARY

Does it make a difference if the default language is set to Maori?

TE TAKA KEEGAN, SALLY JO CUNNINGHAM, KATHERINE DON
University of Waikato
Hamilton, New Zealand

Abstract. In this paper we investigate the effect of default interface language on usage patterns of the Niupepa digital library (a collection of historic Māori language newspapers), by switching the default interface language between Māori and English in alternate weeks. Transaction analysis of the Niupepa collection logs indicates that changing default language affects the length of user sessions and the number of actions within sessions, and that the English language interface was used most frequently.

1. Background

The Niupepa Collection is a collection of historic Māori newspapers published between 1842 and 1933. It is a large source of historic legacy texts, almost 18,000 newspapers pages that were predominantly written in the Māori language. The newspapers are available in a full text and in two facsimile forms: a low resolution image that downloads quickly for previewing, and a high resolution image that is easier to read. The collection is currently served by the Greenstone software[[i]] of the New Zealand Digital Library (NZDL) at: www.nzdl.org/niupepa. For a comprehensive explanation regarding the process of delivering the Niupepa on the Web, see [[ii]].

70% of the documents are written in Māori, 27% are written bilingually in both English and Māori, and 3% written in English only. The collection is a rich source of Māori language texts in an environment where there is a dearth of Māori language resources. The default language of the collection is set to Māori.

Figures released after the New Zealand 2001 census showed approximately 15% (526,281) of New Zealanders to be Māori. Of these only 9% of the adult population has an oral or written proficiency in the Māori language. By setting the default language of the Niupepa Collection to Māori we are clearly going against the preferences of the majority of potential users. Research undertaken by Jones et al [[iii]] suggests that users of a digital library system rarely amend the default settings for options with regard to query types and result displays.

The question that this research seeks to find answers to is what differences will occur in usage statistics and user behaviour if the default language of the interface is alternated between Māori and English.

2. Methodology

2.1. GATHERING THE DATA

The NZDL website (www.nzdl.org) makes available over 40 different collections in various formats. All user activity is logged. Every access or 'hit' is recorded along with information such as the page requested, the language used in the interface, the time of the request, the type of request, the previous action, the ip address of the requestor and the various preferences that are set.

The NZDL site is mirrored with a site located at the University of Lethbridge in Alberta, Canada. The New Zealand site, located at the University of Waikato serves the collection to Web requests from within New Zealand. The Lethbridge mirror site is responsible for serving the collection to Web requests from outside of New Zealand. The data collected in this analysis is from the University of Waikato site only, as thus only reflects usage characteristics that are happening within New Zealand.

We chose to analyse a four week time period running from 8.40am Monday 5 July 2004 to 8:40am Monday 2 August 2004. In the first week we changed the Niupepa default language setting to English. The second week we changed it back to Māori. The third week it was in English and the fourth week we changed it back to Māori again. The raw log file was collected for this time period and the hits relating to the Niupepa collection were extracted.

2.2. CATEGORISING THE INFORMATION

In the log file recorded over the four week period we noticed that the interface language that was being used was English, Māori or an undefined language. The undefined language setting appeared when the language argument was set to either a blank or the number 20. This unusual setting appeared to be either a result of a cookie conflict, or the typing of a bracket character in a search query. It only occurred in a very small percentage of the hits (1.6%) and so consequently these hits were removed from the log file to be analysed.

The subsequent log file was categorised into three further log files. There were the total hits that were recorded over the time period i.e. weeks 1 to 4. There were the hits that occurred when the default language of the website was set to English i.e. weeks 1 and 3, which will be referred to in this paper as the EN weeks. The third category is the hits that occurred when the default language of the website was set to Māori i.e. weeks 2 and 4, which will be referred to in this paper as the MI weeks.

2.3. DEFINING SESSIONS

To further analyse what the users of the web site were doing the hits recorded in the log files were grouped into sessions. Cookies were used to define these sessions. When users connect to the website a cookie is created on their machines which hold information that includes the ip address of the machine connecting and the time that the cookie was created. This information is recorded as an argument, the z argument, and with each hit the z argument is displayed. To group hits by sessions we simply grouped hits that had similar z arguments that occurred with in a given time frame. This time frame was set to 30 minutes, the common minimum time frame setting used in web session analysis.

Once the log file of hits was grouped into sessions, the types of sessions were then arranged based on the length of the session and what the user accessed. This gave three types of sessions:

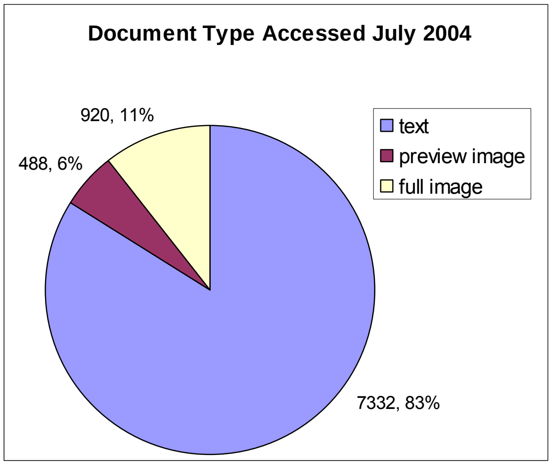
- Single hit sessions: these could be sessions where the user has come into the site and decided not to look any further, or they could be users that have not enabled cookies.
- Exploratory only sessions: multi-hit sessions where the user only accessed the home page, the help page, and/or the preferences page. No documents in the collection were accessed and no searches were undertaken.
- Extended sessions: multi-hit sessions where queries were undertaken and/or documents of the Niupepa collection where accessed.

The extended session are the sessions that we are most interested in as these give an indication on what active users of the collection are doing. We then looked at how many users were using the collection with the interface language set only to English, how many were using the collection with the interface language set only to Māori, and how many users were switching the interface language. The results were analysed for the weeks that the default language was set to English (EN weeks) and then again for the weeks that the default setting was set to Māori (MI weeks).

3. Results Summary

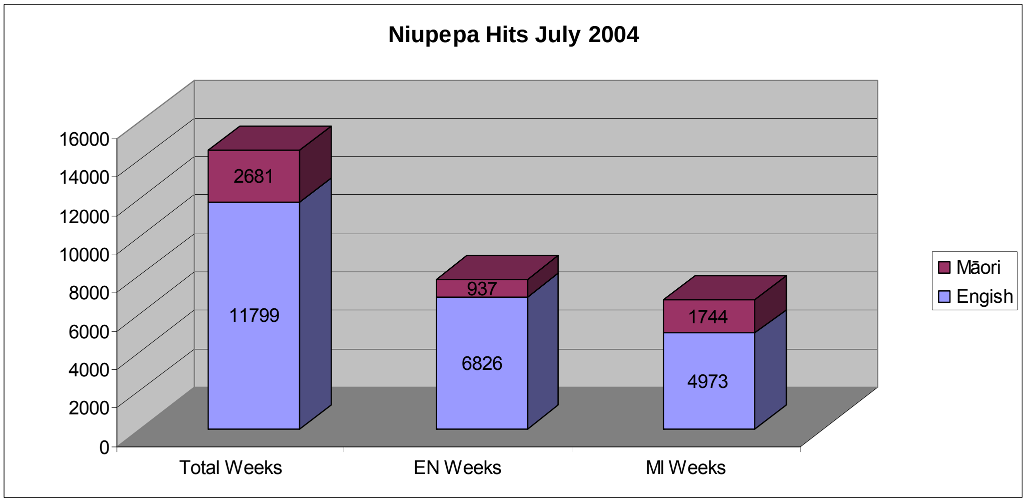
3.1. GENERAL RESULTS

Over the four week period beginning July 5th 2004 and ending August 2nd 2004 the Niupepa collection of the NZDL recorded 14,490 hits. There were 2,360 queries submitted and 8,740 documents from the collection viewed. Of the documents viewed 332 pages (83%) were text files, 920 pages (11%) were high resolution image files and the remaining 488 (6%) were low resolution preview image files. Graph 1 shows the document types that were accessed in this 4 week period.



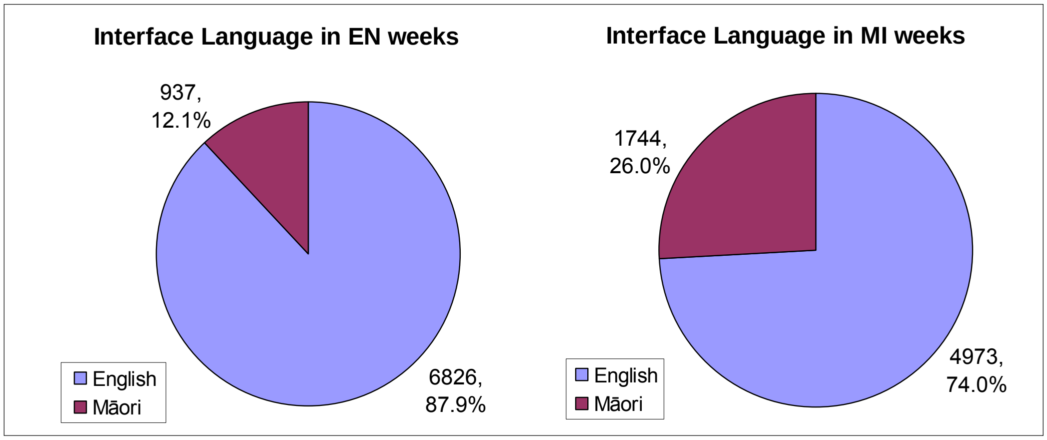
Graph 1: Document Types Accessed in July 2004

With the undefined language hits removed the collection was accessed with an interface language setting of either English or Māori. 81.5% (11,799 hits) were with the interface language set to English and 18.5% (2681 hits) were with the interface language set to Māori. There were more hits in the weeks when the default interface language was set English (EN weeks) with a total of 7763 hits, as opposed to a total of 6717 hits that occurred when the default interface language was set to Māori (MI weeks). The break down of these totals is shown in *Graph 2*.



Graph 2: Overall Language hits for July 2004

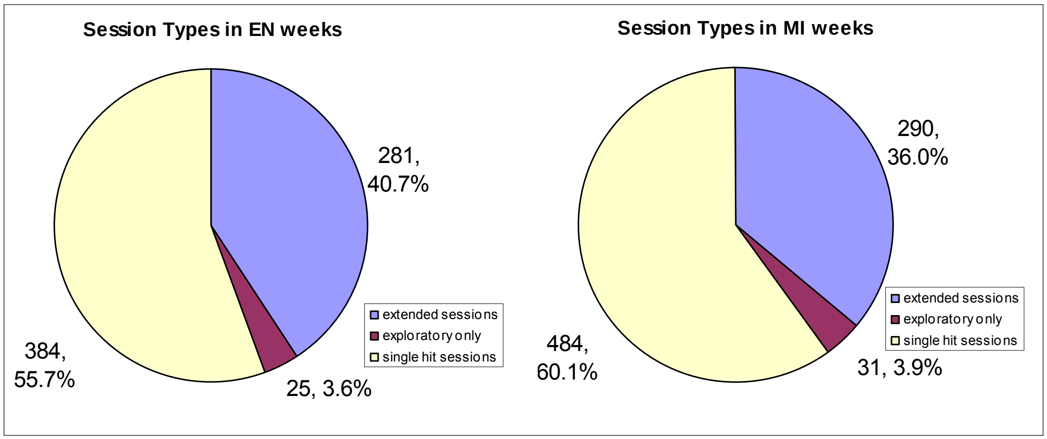
In the EN weeks English was used 87.9% (6826 hits) while Māori was used the other 21.1% (1744 hits). In the MI weeks English was used 74.0% (4973 hits) while Māori was used the other 26.0% (1744 hits). A graphical display of these figures can be seen in *Graph 3*.



Graph 3: Interface Language Usage with different default language settings

3.2. SESSION RESULTS

The hits were separated into sessions as defined in the previous chapter and then the sessions were divided into types of sessions. There were 1495 sessions over the 4 week period; 868 (58.1%) were single sessions, 571 (38.2%) were extended sessions and 56 (3.7%) were page only sessions. Changing the default language of the website did not seem to alter significantly the ratio of session type. In the EN weeks 55.7% of the sessions were single hit sessions and 40.7% were extended sessions. In the MI weeks 60.1% of the sessions were single hit sessions and 36.0% were extended sessions. These are shown in *Graph 4*.



Graph 4: Types of Sessions with different default language settings

From these different types of sessions, perhaps the most useful is the extended sessions. These are users that accessed pages of the Niupepa Collection and/or undertook the entering of search terms. As the accessing of pages from the actual collection and entering of the search terms is quite an interactive process it is unlikely to be an activity that is undertaken by a web robots.

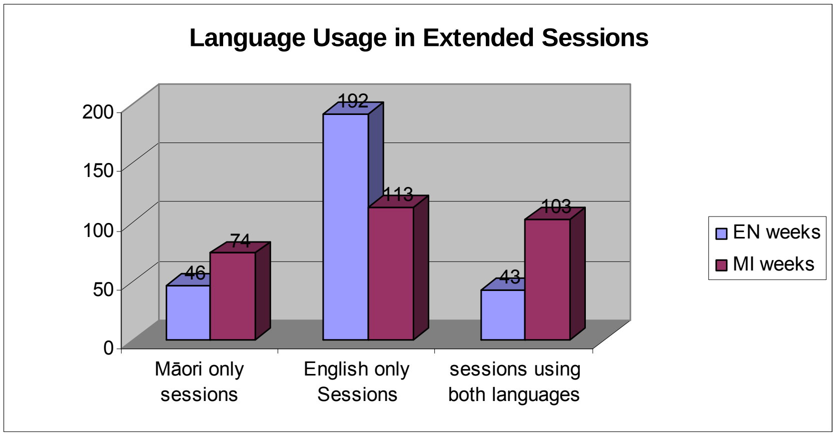
The extended sessions that occurred in the EN weeks appeared to be longer and to involve more hits than in the MI weeks. Table 1 shows that the average session length is 19.2 minutes in the EN weeks and this drops to 16.4 minutes in the MI weeks. The average number of hits is 26.1 in the EN weeks and only 21.3 hits in the MI weeks. There also seems to be a greater dispersion in the times of the sessions and in the number of the hits of the sessions across the two different language defaults as shown by the variance in the respective standard deviations.

	Count	Mean	Median	Mode	Lowest	Highest	Std. dev.
Time (mins) EN	281	19.2	6	0	0	205	31.36
Time (mins) MI	290	16.4	7	1	0	157	23.58
Number of Hits EN	281	26.1	11	3	2	286	37.65
Number of Hits MI	290	21.3	13.5	5	2	136	23.46

Table 1: Statistical Comparisons of Extended Sessions with different default languages

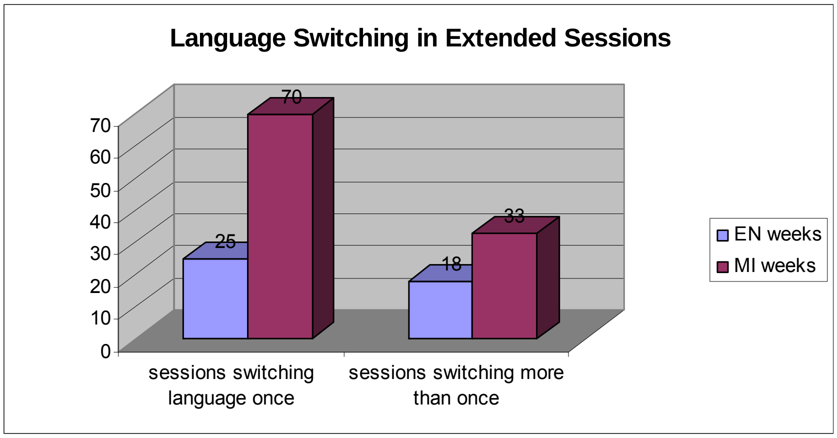
3.3. LANGUAGE USAGE AND LANGUAGE SWITCHING IN EXTENDED SESSIONS

It was perhaps not surprising to note that the usage of a particular language increased when the default language of the interface was set to that language. *Graph 5* shows the number of extended sessions where the interface language remained in Māori (74) is higher in the MI weeks than it is in EN weeks (46). Also the number of extended sessions where the interface language remained in English (192) is higher in the EN weeks than it is in the MI weeks (113). The results also show that language switching occurred more than twice as often in the MI weeks (103) than in the EN weeks (43).



Graph 5: Language Usage in Extended Sessions

A single language switch occurred more often in the MI weeks (70) than it did in the EN weeks (25). This is shown in *Graph 6* which also shows that multiple language switching also occurred in more sessions (33) in the MI weeks than it did in the sessions (18) of the EN weeks. It should be noted that these numbers are small and more data needs to be collected before we can draw some conclusions on language switching.



Graph 6: Language Switching in Extended Sessions

4. Conclusions

Despite the fact that the bulk of the Niupepa documents are in Māori, users of the collection showed a strong preference for the English version of the collection interface. The reasons underlying this preference require further investigation. It was noted that the usage of Māori language version of the collection doubled when the default interface language was set to Māori.

The default interface language affects the usage patterns of the collection: users during Māori language default weeks have shorter sessions and fewer actions within sessions. Additionally, a slightly higher percentage of single hit sessions and brief, exploratory sessions was noted during Māori language default weeks.

[i] for information on the Greenstone software see: Witten, I H & Bainbridge, D. (2002). *How to Build a Digital Library*. Morgan Kaufmann. San Francisco, CA.

[ii] Apperley M D, Keegan 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.

[iii] Jones S., Cunningham S.J., McNab R.J. and Boddie S. (2000) A transaction log analysis of a digital library in *International Journal on Digital Libraries 3*(2) 152-169.