View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Research Commons@Waikato

Language Preference in a Bi-language Digital Library

Te Taka Keegan
University of Waikato
Hamilton
New Zealand
64 7 838 4420
tetaka@cs.waikato.ac.nz

Sally Jo Cunningham
University of Waikato
Hamilton
New Zealand
64 7 838 4420
sallyjo@cs.waikato.ac.nz

ABSTRACT
This paper examines user choice of interface language in a bi-language digital library (English and Māori, the language of the indigenous people of New Zealand). The majority of collection documents are in Māori, and the interface is available in both Māori and English. Log analysis shows three categories of preference for interface language: primarily English, primarily Māori, and bilingual (switching back and forth between the two).

1. INTRODUCTION
As digital libraries increase in number, content, and potential user base, interest has grown in 'multilingual' or 'multi-language' collections—that is, digital libraries in which the collection documents and the collection interface include more than one language. Research in multilingual/multi-language digital libraries and web-based document collections has primarily focused on fundamental implementation issues and functionality [3], principles for design [2], and small-scale usability tests [3]; at present no analysis exists of how these systems are used, or how the presence of more than one language in a digital library affects user interactions—presumably because multilingual/multi-language digital libraries are only recently moving from research lab prototypes to fielded systems, and few have built up a significant usage history.

This paper describes the application of log analysis to examine interface language preference in a bi-language (English/Māori) digital library—the Niupepa Collection (Section 2). Web log data was collected for a year (Section 3), and log analysis indicates three categories of interface language preferences: English, Māori, and 'bilingual' (Section 4). A fine-grained analysis of activities within user sessions indicates different patterns of document access and information gathering strategy between these three categories (Section 5).

2. NIUPEPA COLLECTION
Niupepa (www.nzdl.org/niupepa) is a collection of historic Māori newspapers published between 1842 and 1933 [1]. It is a significant source of historic texts of the indigenous people of New Zealand—just under 18,000 newspaper pages, covering 40 titles. The collection is implemented using the Greenstone digital library software (www.greenstone.org). Keyword searching is supported, and users can browse the collection by newspaper title and chronologically, by issue publication date. 70% of the documents are in Māori, 27% are in both Māori and English, and 3% in English only. The default interface language of the collection is Māori, and an English version of the interface can be selected at any point when interacting with the collection.

3. DATA COLLECTION
All user activity in the Niupepa Collection is automatically logged. The analysis described in this paper is based on a log of local New Zealand usage of the Niupepa from 1 January to 31 December 2004—a total 187,215 hits. This raw data was then further filtered to remove known web robot activity (338 hits), hits where the IP address was not defined (495), hits from the local research team (1565), and hits where the interface language is undefined (3578). The resulting filtered Niupepa log totaled 181,239 hits, comprised of 145,596 hits (80.3%) where the language of the interface is set to English and 35,643 hits (19.7%) where the language of the interface is Māori.

We isolate sessions through an identifying argument stored in cookies. A simple heuristic was used to categorize user 'sessions': a session is assumed to be a series of hits containing the same identifier, and with no more than a 60 minute lapse between consecutive hits.

4. SESSION ANALYSIS
Examination of session length (measured in both time and number of hits) suggested two initial categories of sessions: exploratory and extended. Exploratory sessions are multi-hit sessions where the user accessed only the home page, the help page, and/or the preferences page. No documents were accessed and no searches were undertaken. Many of these sessions simply involved switching the default interface language from the default Māori to English; presumably these users are simply gaining enough of an overview of the collection to decide that it is not of interest to them. Exploratory sessions are not considered further.

Extended sessions are multi-hit sessions including at least one search, browse, or document access; the log includes a total of 5161 extended sessions. We define the preferred interface language of a session as being the language setting for at least 80% of the hits in a session, where the session does not involve more than two user interface language switches. In Table 1, we see that the sessions do not cleanly divide into Māori and English preference sessions; a significant minority (7%) of the sessions were 'bilingual' in the sense that they involved three or more user interface language switches and/or the interface was set to English and to Māori for at least 20% of the session's hits.

Note that English is the preferred interface for over two thirds of the sessions (68.5%)—an apparently counter-intuitive outcome, given that the vast majority of the Niupepa documents are in Māori, and it would seem reasonable that users would prefer the language of the interface to match the language of the documents. However, consideration of the language demographics of New Zealand offers an explanation. Approximately 14% of New Zealand's population is Māori, and only one in four Māori are able to converse in that language. Historically Māori has been primarily an oral language, so it likely that the percentage of potential Niupepa users who are fluent in written Māori is much smaller than the number of Māori speakers—and so the high proportion of English interface preference sessions is likely to reflect a greater fluency in written English than in written Māori among Niupepa users.

Where within sessions are interface language switches occurring? In the Māori and English sessions switches primarily occurred in the first quarter of a session. In bilingual sessions, the switching of the language of the interface occurred evenly throughout all quarters of the session—an interesting observation, perhaps indicating that the user in bilingual session feels equally comfortable (or uncomfortable!) in both interface languages.

Table 1. Niupepa Collection Session Activity 2004

	English	Māori	Bilingual
sessions:	3548	1267	366
total session %:	68.5%	24.5%	7.1%
page hits:	108479	29055	7845
average (hits):	30.6	22.9	21.4
median (hits):	15	7	9
average (minutes):	25.2	18.7	16.3
avg newspaper page retrievals /session	5.43	15.87	10.82
avg searches/session	4.8	3.4	3.1

5. PROFILE OF SESSION ACTIVITIES BY LANGUAGE PREFERENCE

A fine-grained analysis of session activities shows different patterns of use among the three interface language preferences—specifically, in number of newspaper pages retrieved, choice of document format, and use of searching or browsing as the dominant information seeking strategy.

Māori preference sessions include significantly more newspaper page retrievals per session (average 15.87 newspaper pages) than bilingual (10.82 pages/session) and English preference sessions (5.43 pages/session) (Table 1). Use of the Māori interface is clearly linked to success in locating documents of potential interest—presumably also indicating a higher degree of fluency in reading Māori. When accessing a newspaper page, users can choose to view either an enlarged, high-resolution facsimile of the original page, a smaller, low-resolution (and less readable) facsimile, or the extracted text. Document retrievals in Māori preference sessions included a relatively higher proportion of requests for high-resolution facsimiles (21.5% of Māori session newspaper page retrievals, in comparison with 9.9% of bilingual session newspaper page retrievals, and a mere 5.6% of page retrievals in English sessions). The high-resolution facsimile is intended for on-screen reading—supporting the assumption of greater written Māori fluency in Māori preference sessions.

Browsing was used to a greater extent in bilingual and Māori preference sessions than in English preference sessions: newspaper pages were accessed from one of the browsing interfaces for 28.4% of all newspaper page accesses in bilingual sessions, 26.4% of page accesses in Māori preference sessions, and 15.2% of all page accesses in English preference sessions. Searching occurs more frequently in English preference sessions (average of 4.8 searches per session) than in Māori preference (3.4 searches/session) and bilingual sessions (3.1 searches/ session). The relative preference for searching rather than browsing in English interface sessions is highlighted when we look at the percentage of sessions that did not include searching: 37.2% of bilingual sessions and 35.4% of Māori preference sessions did not include searching, compared with 24.8% of the English preference sessions. Effective browsing requires a greater fluency in Māori than does searching, as browsing is primarily over Māori language newspaper titles and content. Searching allows a less fluent user to focus more tightly on specific documents (that is, those newspaper pages containing the search terms).

6. CONCLUSIONS

Log analysis of the bi-language Niupepa collection clearly indicates three categories of preference for interface language. These different interface language preferences are linked to different patterns of activities within sessions—and so to the potential to offer greater support for each category of user when designing a digital library. In the Niupepa Collection, English preference interface users may benefit from enhanced searching facilities and from better facilities for browsing over English language metadata; these design tactics would allow these users to minimize the need to read Māori text as part of the mechanics of information seeking, as these users narrow their focus to locate documents of potential interest. Māori interface preference users may benefit from enriched Māori language browsing facilities; the current browsing structures are primarily over newspaper series title—giving meager insight into the newspaper contents.

Inevitably log analysis raises as many questions as it answers, as we can only examine user actions and not their intentions or information needs. Of particular interest is the identification of bilingual users; are these users switching interface language to reflect changes in their information need as the session progresses, or are the switches a sign of frustration with the system, or is there another explanation entirely? Further research (including in-depth studies of small groups of users) is indicated to explore the basis for these different behaviors.

7. REFERENCES

[1] Apperley M. D., Keegan T. T., Cunningham S. J., & Witten, I. H. Delivering the Māori newspapers on the Internet. In *Rere Atu Taku Manu! Discovering History Language and Politics In The Māori Language Newspapers*. Auckland University Press. (2002) 211-236.

[2] Huang, S., & Tilley, S. Issues of content and structure for a multilingual web site. In *Proceedings of SIGDOC '01* (2001), 103-110).

[3] Perufini, S., McDevitt, K., Richardson, R., Perez-Quinones, M., Shen, R., Ramakrishnan, R., Williams, C., & Fox, E.A. Enhancing usability in CITIDEL: multimodal, multilingual, and interactive visualization interfaces. In *Proceedings of JCDL '04* (2004), 315-324.