ANALYSIS OF THE NEW ZEALAND ENGLISH AND MĀORI ON-LINE TRANSLATOR

Mark Laws, Richard Kilgour and Michael Watts
Knowledge Engineering Laboratory, Department of Information Science,
University of Otago, P.O.Box 56 Dunedin, New Zealand.
Email: maaka@kel.otago.ac.nz, richard@kel.otago.ac.nz, mike@kel.otago.ac.nz

Abstract - The English and Māori word translator *ngā aho whakamāori-ā-tuhi* was designed to provide single head-word translations to on-line web users. There are over 13,000 words all based on traditional text sources, derived because of their high frequency used within each of the respective languages. The translator has been operational for well over a year now, and it has had the highest web traffic usage in the Department of Information Science. Two log files were generated to record domain hits and language translations, both provided the up-to-date data for analysis contained in this paper.

1. INTRODUCTION

The need for an effective method of managing the increasing speech data repositories called for the construction of a relational database management system, containing various speech and language structures with phonological, lexical, semantic, syntactic and morphological formations. The Management of Otago Speech Environment (MOOSE) database [1] is the resulting tool. It is a bilingual model that allows full interaction with an English and Māori computer-based text/speech architecture [2]. Because the database houses a lexicon containing vocabularies in both languages [3, 4], initially a simple query was written to provide a basic head word translation. Further development extended the system into a separate functional database for the Windows environment, using Microsoft Access 97 to generate the language tables, queries and an interface to meet the bilingual requirements. An indexing structure was incorporated to match a word from one language set to corresponding words in the other language set. Thus, a 'one to many' translation method was achieved. The translations from the target language are words listed in order of their search result, there are no grammatical conventions associated with the words or any identifiers to show different meaning or application [5]. A feature of this translator allows some of the Māori searched words to be linked to over 500 pronunciations in common use today [3].

2. WEB PAGE DEVELOPMENT

To widely test the translator with other users from different domains, a version on our internet server was developed (http://kel.otago.ac.nz/translator/). The technical specifications of the server 'kel' are a Pentium II 266MHz processor, 64 MB RAM, running Red Hat Linux version 6.0.

The translator is based entirely upon open source software. Open source was chosen for several reasons:

- low cost to initially create the system
- high reliability
- system stability under heavy loads
- multi-platform support–the option to port the system to other machines and operating systems with minimal modification.

The Apache HTTP server (http://www.apache.org) is the most widely used web server in the world. Apache has consistently demonstrated levels of reliability, security and flexibility far in excess of its competitors.

The translator functionality is driven by version 3 of the PHP scripting language (http://www.php.net). PHP has several advantages over other scripting languages:

- native database support
  - eases the programming burden
  - adding efficiency of execution
- automatic parsing of arguments to the requested file
- an embedded Apache module
  - improves overall security.

PHP scripts can be embedded into HTML files, which makes such things as a bilingual interface possible. By passing the desired language as an argument in the URL, the script can determine which language to use for the text of that page.

The database PostgreSQL version 6.5.1 (http://www.postgresql.org/) was chosen due to it's improved database management functions, specifically, views and triggers.

Currently 34 known web pages have links back to the translator site. The majority of these pages are either collections of links to online dictionaries or information about Māori and New Zealand culture. The remainder are educational resource sites and personal web pages.

3. A BILINGUAL SYSTEM

A true bilingual system gives both languages equal status [6]. Consequently, the option to switch the interface between the two languages was an important feature to consider. Users can therefore read all the instructions, messages, dialogues and comments in their preferred language.

Table 1: Bilingual text on the translators web interfaces.

Instructions:
Enter a single word to translate...
He kupu tētahi kia whakamāori...
Sorry, the English Word "" is not in the database
Ka aroha, kaua tēnei kupu Pākehā "" ki roto i te pātengi raraunga
Messages:
English Word: <word>
He Kupu Pākehā: <word>
Click on the word link to play the pronunciation of that word.
He kupu pāwhiri kia whakahoki me te whakahua i te reo Māori
Button Dialogues:
Translate to Māori
Kia whakamāori
Translate another word
Whakamāoritia he kupu tētahi anō
Comments:
This page is maintained by Michael Watts
Kua whakaū ki tēnei wharangi ā Michael Watts

The bilingual interface promotes the system to a wider audience [7], it also recognises that Māori is the official national language of New Zealand. The interface allows users to enter the word (to be translated) into the search box, then choose the appropriate translation method (e.g. select the "Translate to Māori" button). The resulting translation(s) are then presented on another page.

4. WEB PAGE ANALYSIS

According to Lawrence and Giles [8] scientific and educational web pages are only 6% of the total web content. On-line translators fall within these two categories, therefore they would only provide a very small portion of that 6% range. This papers analysis also reflects the small number of web users who have accessed this translator over the short time period.

The first log file was generated by the Apache server, it records all access traffic and therefore the number of hits on the translators web page. This has been active since August 1998. The steady increase in hits was interrupted in August 1999, which reflects the submission of the translators URL to various national and international web resource sites. For example, links were added to pages that made reference to on-line translators.

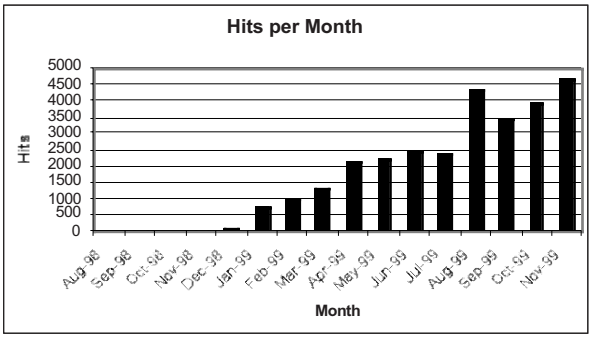

Fig 1. Translation page hits.

The second log file was commissioned during March 1999 to record the database transactions. The PHP3 translation script was modified to log all word submissions, so approximately eight months of on-line translations from the web page were analyzed. Note that translations from machines within the University of Otago were excluded, as the vast majority of these were performed for in-house testing purposes only.

4.1 Domains

The first metric filtered the number of domains from the client users. 9,703 domain hits were available, given that over half the domains were not registered due to those clients machines not being associated with a DNS entry. New Zealand (nz) accounts for over half the domain hits, which is expected given the high interest by local language users. The other main domains represented show a similar pattern to other analyses undertaken in New Zealand on a similar on-line English-Māori lexical database[5].

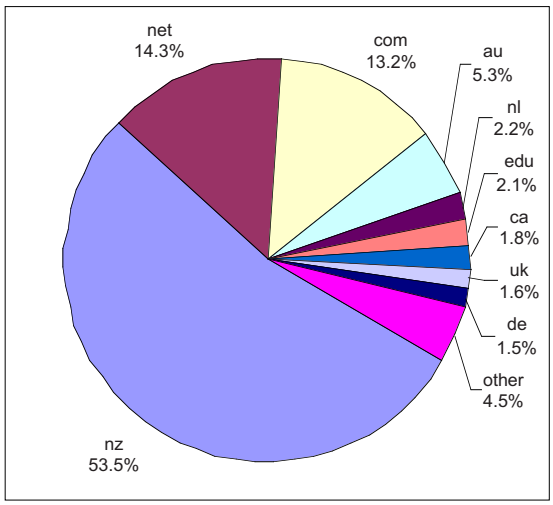

Fig 2. Translation page hits by domain.

The next series of metrics focused on the number of words successfully and unsuccessfully translated, the mis-hits and user errors. In the current time-frame of analysis, only a small number of new words were added to the database from either language. Additions and other improvements will be implemented soon after the results of this paper are presented for discussion.

4.2 English Words

Of the 13,530 English words submitted, 9,210 (68.1%) were successful in translation. Many of those words were requested a number of times each, given that only 4,517 different English words were actually retrieved from the database. The most common 100 words accounted for 30.5% of the total number of

requests submitted. The ten most common requests are shown in Table 2. The size of the English vocabulary is restricted to 6,880 words, so the majority of words that were not successfully translated were not in the database.

Table 2: The most common English word requests.

Rank:	Word Request:	Translation:	Hits:
1	hello	*kia ora*	324
2	love	*aroha*	274
3	goodbye	*e noho rā*	125
4	I	*ahau*	123
5	you	*koe*	122
6	house	*whare*	107
7	welcome	*haramai*	84
8	dog	*kurī*	75
9	good	*pai*	72
10	friend	*hoa*	72

Other common categories of mis-hits include: short phrases (e.g. "I love you", "good morning"), Māori words or place names (e.g. *kiwi*, *Tauranga*) being submitted by users who choose the incorrect translation method (i.e. they selected the wrong translation button), and profanity–given that only the most commonly used written words in both languages are used means there are currently no profane entries.

4.3 Māori Words

Requests for Māori word translations were fewer than English words. Here, 8,729 words were submitted of which 3,719 (42.6%) were successfully translated. The most common reason for not finding a translation was the incorrect use of the macron characters. For example, *pakeha* was requested four times more often than *pākehā*, although only the latter appears as the correct spelling in the database.

The other most common reasons for mis-hits were users submitting English words and using the English translation method instead of the Māori option (words such as "love" and "hello" were requested several times in this way). Multiple word requests (e.g. phrases or sentences) also account for several mis-hits, although substantially less than English word requests.

Table 3: The most common Māori word requests.

Rank:	Word Request:	Translation:	Hits:
1	*te*	the	78
2	*kia*	let	71
3	*kia ora*	hello	64
4	*aroha*	love	62
5	*ka* (mis-hit)	...	60
6	*Māori*	native	52
7	*nui*	big	46
8	*kai*	food	46
9	*pakeha* (misspelt)	european	44
10	*ora*	life	43

The most common mis-hits and unsuccessful translations for both languages are shown in Table 4. This indicates that in the first instance the database has insufficient words (and phrases) to provide translations for the requested words. Secondly, the user is applying the incorrect translation method to their word requests. Finally, the user had spelt the words incorrectly–especially for Māori.

Table 4: Most unsuccessful words from each language. Note the * indicates that the word has been misspelt.

English Words:	Mis-hits:
thank you	55
thanks	54
thank	48
fuck	45
sex	41
new zealand	41
i love you	37
good morning	27
thankyou *	26
greetings	23

Māori Words:	Mis-hits:
ka *	60
pakeha *	44
nga *	37
arohanui	31
aotearoa	30
ra *	24
whanau *	23
whaka *	23
tane *	23
hapu *	23

4.4 Bilingual Interface

This next section looks at the bilingual interface in terms of how many users actually switched the view to read all the text in Māori, and once in that mode how many times they actually used the translator. Because the bilingual aspect of the translator was fully implemented late in the on-line process, only three complete months of log files exist for analysis. These results can still be looked at as an early indication to what extent this facility has allowed users choice. It may not currently be evident in the overall picture, but it is one of the important reasons why the translator is being accessed by a small section of users.

The number of times the bilingual interface was selected totaled 1,060, and the number of times the interface was used for translations was 527 (49.7%). A further breakdown showed that each of the three months had an even distribution of interface and translation totals. Therefore, this seemed to indicate a regular rate of use. It also appears that about half of the time the interface was only switched to Māori for perusal purposes by those users .

4.5 Māori Speech

The final analysis was based on comparing the frequency of the 500 Māori pronunciation sound files to their equivalent Māori translation results requested in either language. For example, the *aroha* file was accessed 163 times, it was presented in the users translation results interface 336 times, English (274), Māori (62), therefore the access frequency of this file was 48.5%. The pronunciation sound files were accessed a total of 3,327 times, which includes 408 of the 500 files available. The following table shows the ten most commonly used pronunciation files and

their association with the English and Māori search results.

Table 5: The most common Māori Pronunciation Files (MPF) associated with the word request rankings in both languages.

MPF:	Hits:	English:	Māori:	Frequency:
aroha	163	274	62	48.5%
whare	76	114	19	57.1%
haramai	69	109	3	61.6%
haka	62	17	27	140.9%
koe	61	122	27	40.9%
ahau	49	166	16	26.9%
kai	47	83	46	36.4%
tangata	43	86	21	40.1%
kurī	43	75	12	49.4%
hoa	41	51	11	66.1%

Table 5 is interesting as it demonstrates that, on average, for over half the translations in the results pages the users are playing the pronunciation files. With the exception of *haka*, the English requests show a similar pattern to the number of times the files are accessed. Further analysis into why *haka* is played more times than it is actually requested showed that this file was accessed multiple times per translations, it was also directly accessed by search engines, and users had also bookmarked the file location. A conclusion to this anomaly suggests that because this word is associated with the New Zealand All Black Rugby team's legendary pre-match *haka* (dance), that its popularity is disproportional by comparison to the other most common Māori words.

5. SUMMARY

It is clear that the results of this analysis presents an all important step towards providing a comprehensive on-line bilingual text and speech based resource.

Based on the current results, future translators will provide the user with speech synthesis tools to compliment the text entry and reporting features that already exist. This functionality will allow the translated English and Māori words to be generated into speech using current diphone database techniques with a multi-lingual speech synthesiser called 'MBROLA' and a 'text-to-speech' system called 'Festival' [9]. Current experiments on the bilingual synthesiser will be released soon. The bilingual text entry facilities will also be supported with the 'ISpell' [10] spelling checker, which again will assist the correct submission of word requests. All three systems are open source supported.

ACKNOWLEDGEMENTS

Support from CBIIS UOO-606 and UOO-808 funded by the New Zealand Foundation for Research Science and Technology (NZFRST). Other contributions from the University of Otago Post Graduate Scholarships and the NZFRST Tūāpapa Pūtaiao Māori Fellowship scheme. Dr Godfrey Pōhatu and Ms Lorraine Johnston for assisting with the bilingual interface translations. Mr Peter Keegan and Mr Te Taka Keegan for providing information on the Māori language vocabulary.

REFERENCES

[1] Laws, M., Kilgour, R. (1998) *MOOSE: Management Of Otago Speech Environment*. The 5th International Conference on Spoken Language Processing, Sydney Convention and Exhibition Centre, Darling Harbour.

[2] Kasabov, N., Kozma, R., Kilgour, R., Laws, M., Taylor, J., Watts, M., Gray, A. (1999). *A Hybrid connectionist-based methods and systems for speech data analysis and phoneme-based speech recognition*. In: Neuro-Fuzzy Techniques for Intelligent Information Processing, N. Kasabov and R. Kozma, (eds.), Heidelberg Physica Verlag.

[3] Benton, R. A., Tumoana, H., Robb, A. (1982*) Ko Ngā Kupu Pū Noa o Te Reo Māori: The First Basic Maori Word List*. New Zealand Council for Educational Research, Wellington.

[4] Bauer, L. (1994) *Introducing the Wellington Corpus of Written New Zealand English*, Te Reo, Journal of the Linguistic Society of New Zealand, Volume 37, University of Auckland. (http://www.vuw.ac.nz/lals/corpra.htm)

[5] Keegan, P. (1997) *Kimikupu Hou Māori Lexical Database on the Web: Reflections and Possible Future Directions*. In Proceedings NAMMSAT Conference, October 1997, Massey University, Palmerston North.

[6] Laws, M. (1998) *A Bilingual Speech Interface for New Zealand English to Māori*. Unpublished M.Sc. thesis, University of Otago.

[7] Crystal, D. (1992) *The Cambridge Encyclopedia of Language*, Cambridge University Press, Cambridge.

[8] Lawrence, S., Giles, C. L. (1999) *Accessibility and Distribution of Information on the Web*, Nature, 400, 107-109. (http://www.wwwmetrics.com/)

[9] Laws, M. (1999) *A Bilingual Information System - The Computational Linguistic Engineering of English and Māori for Speech Perception and Generation*. Proceedings of the ICONIP/ANZIIS/ANNES'99 International Workshop "future Directions for Intelligent Systems and Information Sciences". pp 75-78.

[10] ISPELL V3.1 by G. Kuenning, *et al*. (http://www.delorie.com/gnu/docs/ispell/)