Minority Languages & Translation Technologies Case Study: Te Reo Māori & Google Translator Toolkit

Introduction

This paper describes a case study where the Google Translator Toolkit (GTT) was used to undertake a large translation task involving a minority language. The task involved the translation of 50,000 interface terms of the Moodle learning management system into te reo Māori (the Māori language).

The paper begins by describing how some minority languages are not in an environment where technology can be easily used for translations. It then suggests that te reo Māori however, is in a position to use technology and suggests some technologies that are suitable. The paper then briefly describes the GTT and the translation task that this tool was used for. The translators' feedback on the use of this technology in this environment is summarised and it appears that the GTT is suitable to be used by minority language translators.

Minority Languages and Technology

It is widely accepted by linguists that there are currently about 7,000 languages of the world (Austin & Salback, 2011, p1) and the weighting and distribution of those languages are heavily skewed. There are a small number of languages that are spoken by a large number of speakers and a large number of languages that are spoken by a small number of speakers. Only 5% (347) of the world's languages have 1 million or more speakers, but this covers 94% of the world's population. The remaining 95% of the languages (6565) are spoken by only 6% of the world's populations (Lewis, 2009). Simple economics suggest that the languages with the largest number of speakers will always get the largest support in terms of the creation of language technologies and resources.

The smaller languages, which we will term in this paper as minority languages, suffer from a dearth of resources. This paucity of resources can be grouped into 6 main areas.

1. **Lack of people**: – many minority languages suffer from a lack of people to assist with the creation of language technologies. There could be a lack of actual speakers of the language, a lack of people knowledgeable about the language, a lack of people willing to use modern technologies, a lack of people willing to create and take responsibility for the technologies, a lack of people with knowledge on how to create language technologies and even a lack of people that believe there is merit in placing language resources in an electronic environment.
2. **Lack of content**: – many minority languages lack a significant amount of digital content that is necessary to build language technologies. Some languages do not have a written form at all, while some languages have written forms that are not stored in an electronic format and thus may require significant effort to digitise the language resources.
3. **Lack of trust**: – many minority languages treat their language resources as sacred treasures that need special care and protection. There is a concern that placing language resources and consequently cultural knowledge in electronic environments can open them up to abuse and degradation. There are also concerns about privacy and intellectual property rights.
4. **Lack of finance**: – creating language technologies cost money and time. Minority language communities are invariably communities operating in the lower ends of the economic scales and subsequently are the ones least likely to have money to spend on creating language resources. They are also the communities who struggle the most with health, education, housing and other social issues, so are the communities least likely to have the time to investigate, strategize and implement programmes that utilise language technologies.
5. **Lack of tribal unity**: – some minority language communities, do not agree on a single form of language, or a single written form of language, or a specific authority to make decisions about how a language should be represented electronically. Furthermore, minority language communities under duress due to inter tribal politics, or even inter tribal warfare face extra hardships and are unlikely to be in a position to consider language technologies.
6. **Lack of government support**: – some minority languages receive little in the way of support or recognition from government authorities. Some government authorities actively seek to repress minority languages e.g. the Kurds in Turkey (Nettle & Romaine, 2000, pp145-146), and the Siraya people in Taiwan [1].

Consequently many minority languages may not be in an environment where language technologies can serve a meaningful purpose. However the authors believe that this is not the case with regard to te reo Māori, and that this language can in fact benefit greatly by the uptake of language technologies.

Te reo Māori

Te reo Māori is an Eastern Polynesian language spoken predominantly in Aotearoa New Zealand and all most exclusively by Māori people. Estimates on the number of speakers and level of te reo Māori speakers vary, but according to the 2006 census (Statistics NZ, 2006) there were just over 165,000 people who could hold a conversation in Māori about everyday things. This represents almost 1/4 (23.7%) of Māori and a little over 1/20 (4.2%) of the Aotearoa New Zealand population. There are very few (if any) monolingual speakers of Māori, with all having an ability to converse in the predominant English language.

Having 165,000 speakers means that te reo Māori is ranked 322 in terms of number of speakers (one below the Navajo language) (Lewis 2009). Given that the Internet penetration ratio in Aotearoa is 80.5% [2], a figure likely to be lower for Māori who suffer from lower income and likely to be even lower for te reo Māori speakers who are mostly senior citizens or younger children, a conservative figure could suggest maybe 100,000 users of the Internet are te reo Māori speakers. 100,000 possible Internet users in a pool of 6.7 billion Internet users is just 0.002%, thus te reo Māori on the Internet is indeed a small minority!

There are some significant electronic resources available in te reo Māori that can be used to assist with the creation of language tools. The bible has been translated into Māori, some large user interfaces have been translated (e.g. MS Office, Windows, Google interface), there is a legacy newspaper archive in Māori, on-line dictionaries, websites and books that are available on-line (in Māori). The resources are held by various institutes and many come under the control of Te Taura Whiri [3].

Te Taura Whiri has been instrumental in a standard te reo Māori orthography and the creation of new terminology. It has also been aware of the role of new technologies in promoting te reo Māori development and has made available electronic resources from its bilingual website (Keegan, Keegan & Laws, 2011, pp2-3).

A recent government review of te reo Māori sector and te reo Māori strategy listed among its recommendations to “embrace technology as modern tools in the revitalisation” (Te Puni Kōkiri, 2011, p23). This suggests that the NZ government does recognise the importance of language technologies to the survival and growth of te reo Māori.

Given that there is a body of te reo Māori speakers, there is a body of electronic content (which in turn implies some level of trust in electronic resources), there is a unity of tribal resources (through Te Taura Whiri) and there is government support for te reo Māori, then it suggests te reo Māori is suitable for use with language technology. A key ingredient still lacking is funding for te reo Māori technology; to which an obvious solution is to use language technologies that are available at no cost.

Language Technologies that can assist Translators

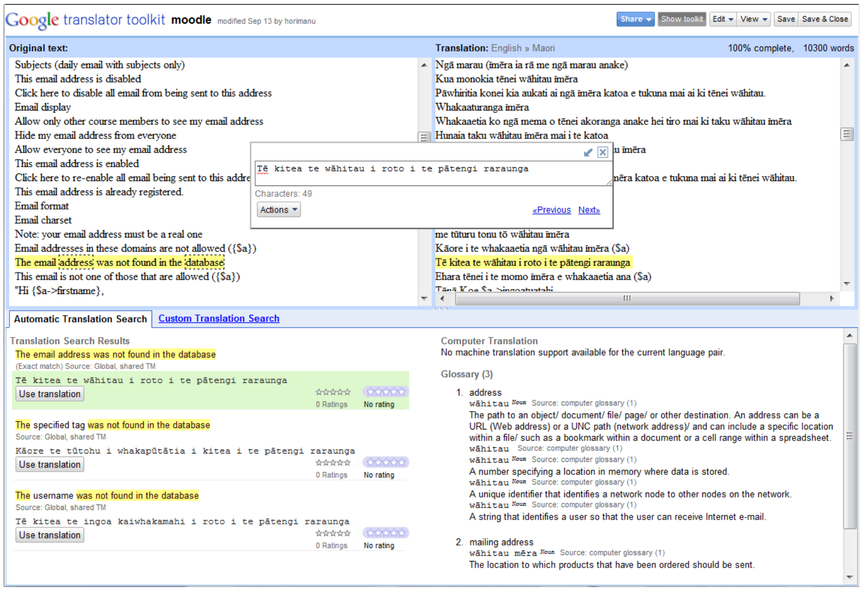
There are many language technologies that can assist minority languages however the purpose of this paper is to discuss one type of technology that is used to assist language translators. This type of tool is commonly referred to as a Computer Assisted Translations (CAT) tool. CAT tools are computer software that is used to assist a human translator to translate texts between languages. This software can exist in many forms and can range from: spell & grammar checkers, to terminology management systems (including dictionaries), to translation memory systems, to machine translation systems, to full translation and project management systems.

The software under study in this paper is classified as a translation memory (TM) system. This technology has caught the interest of the Welsh Language Board who have described it in the following manner:

> Translation memory software is a piece of software that keeps a record of previous translations. When a new document is translated using the software the software will search its memory for components that can be used to translate the new piece of translation. These components include terms or complete phrases... Translation memory software analyzes how much of the content of the new document matches the contents of the translation memory.
> The use of translation memory software can lead to financial savings for individual translators as shown above. It may also be of benefit to larger translation units and companies, as some types of software allow individuals in different offices and organizations to use the same software and work on the same document at the same time.
> … This can help the translator in several ways:
>
> - It leads to consistency in style and terminology
> - It can allow translators to translate more words in less time
> - There is no need to re-translate text that has already been translated.
>
> Translations loaded into the software should be of a high standard in order to increase the likelihood that translations offered by the software will be of high quality. By using the software the translator will then gradually refine the contents of the translation memory. (Welsh Language Board, 2011, p21).

The Google Translator Toolkit

The Translation Memory (TM) system used in this case study is the Google Translator Toolkit [4]. This tool was released by Google in 2009 and is available on-line through a web browser interface at no cost. See *Figure 1*. It has a WYSIWYG (What You See Is What You Get) interface and allows for the sharing of translations in real time amongst any number of translators.


Google translator toolkit moodle modified Sep 13 by horimanu
Share
Show toolkit
Edit
View
Save
Save & Close
Original text:
Translation: English » Maori
100% complete, 10300 words
Characters: 49
Actions
«Previous
Next»
Automatic Translation Search
Custom Translation Search
Use translation
0 Ratings
No rating

Subjects (daily email with subjects only)
This email address is disabled
Click here to disable all email from being sent to this address
Email display
Allow only other course members to see my email address
Hide my email address from everyone
Allow everyone to see my email address
This email address is enabled
Click here to re-enable all email being sent to this addre
This email address is already registered.
Email format
Email charset
Note: your email address must be a real one
Email addresses in these domains are not allowed ({$a})
The email address was not found in the database
This email is not one of those that are allowed ({$a})
"Hi {$a->firstname},

Ngā marau (īmēra ia rā me ngā marau anake)
Kua monokia tēnei wāhitau īmēra
Pāwhiritia konei kia aukati ai ngā īmēra katoa e tukuna mai ai ki tēnei wāhitau.
Whakaaturanga īmēra
Whakaaetia ko ngā mema o tēnei akoranga anake hei tiro mai ki taku wāhitau īmēra
Hunaia taku wāhitau īmēra mai i te katoa
u īmēra
mēra katoa e tukuna mai ai ki tēnei wāhitau.
me tūturu tonu tō wāhitau īmēra
Kāore i te whakaaetia ngā wāhitau īmēra ($a)
Tē kitea te wāhitau i roto i te pātengi raraunga
Ehara tēnei i te momo īmēra e whakaaetia ana ($a)
Tēnā Koe $a->ingoatuatahi

```
Tē kitea te wāhitau i roto i te pātengi raraunga
```

Translation Search Results

The email address was not found in the database
(Exact match) Source: Global, shared TM
Tē kitea te wāhitau i roto i te pātengi raraunga

The specified tag was not found in the database
Source: Global, shared TM
Kāore te tūtohu i whakapūtātia i kitea i te pātengi raraunga

The username was not found in the database
Source: Global, shared TM
Tē kitea te ingoa kaiwhakamahi i roto i te pātengi raraunga

Computer Translation
No machine translation support available for the current language pair.

Glossary (3)

1. address
   wāhitau Noun Source: computer glossary (1)
   The path to an object/ document/ file/ page/ or other destination. An address can be a URL (Web address) or a UNC path (network address)/ and can include a specific location within a file/ such as a bookmark within a document or a cell range within a spreadsheet.
   wāhitau Source: computer glossary (1)
   wāhitau Noun Source: computer glossary (1)
   A number specifying a location in memory where data is stored.
   wāhitau Noun Source: computer glossary (1)
   A unique identifier that identifies a network node to other nodes on the network.
   wāhitau Noun Source: computer glossary (1)
   A string that identifies a user so that the user can receive Internet e-mail.
2. mailing address
   wāhitau mēra Noun Source: computer glossary (1)
   The location to which products that have been ordered should be sent.

Figure 1: The Google Translator Toolkit

The software operates though a number of fundamental steps. First a source document is uploaded into the TM tool. The TM tool then divides the document into translatable segments, usually sentences. Two views of the document are then displayed; the original source document on the left, and the TM segmented version of the document on the right. Segments are presented to the translator in an editing window and resources are also provided to the translator to assist with his translation. These resources include pre-translated segments (returned from TM memory) and translated words (from an associated glossary). The translator considers the suggested translation and glossary words and then decides whether to accept the translation, to enhance it, or to completely re-write it. When complete a click on the 'next' button will move the translator through to the next segment. When all segments have been translated the document is saved in its translated form.

The Translation Task

Introduction

Moodle is an open source, on-line, teaching and learning management system (see www.moodle.org). The University of Waikato has used Moodle as an e-learning platform to deliver and support courses since 2006. Moodle has been translated in many languages; currently 85 languages are available for use with various functions. Moodle's architecture is modular by design and consequently language strings for each language can reside in different areas based on the type of tool or module. Two te reo Māori packs existed for Moodle covering various modules; one that was translated by staff at the Waikato Institute of Technology (WINTEC) and made available in 2005, and another that was translated by the University of Waikato and made available in 2007.

However these language packs had not been updated since their initial release and with the recent upgrade to Moodle 2.0 the University of Waikato's WCEL (Waikato Centre for ELearning) team noticed that only 47% of the user strings were available in te reo Māori. Consequently discussions began in September 2010 to re-visit and re-translate the reo Māori pack for Moodle 2.0.

Requirements

Due to time and financial constraints a decision was made not to translate the complete set of Moodle 2.0 strings into Māori, but instead translate those strings that would be displayed to the core users (i.e. the most common screens displayed to students). Troy Williams and Teresa Gibbison of the WCEL team prioritised and selected 68 of the 130 Moodle 2.0 files of user interface strings that needed to be translated. Two php scripts were written, one for exporting to MS Excel format and the other importing from MS Excel format back to the Moodle format in a php file. For the exporting process Moodle language API's were used to get all strings then export them to excel files based on the module or component name. File sizes varied considerably with the average being close to 840 words. The total source text word count of all 68 files was over 57,000 words. This task required the translation of 35,000-40,000 words of user interface strings.

The material for translation was extracted as source text strings and exported in Excel spreadsheet files for review. In total there were 130 individual Excel spreadsheet files that required review, each containing varying amounts of source text to be translated. The files ranged in size with the smallest containing a mere single source text word, to the largest file exceeding 10,300 words. There was a six month time-frame in which to complete the job, which entailed planning, research, translation and verification.

Translation Issues

Translation work of this nature requires skills above and beyond those possessed by your average translator of te reo Māori. At the very least translators need to have a high level of proficiency with te reo Māori, and be able to appropriately translate new computer jargon. Most of all, they require competency gained from experience using Moodle in order to fully appreciate and understand contexts. Being familiar with the various Moodle functions and activities displayed in the source language, provides the translator insight from the perspective of a typical Moodle user. All of the translators that assisted in the translation of the Moodle 2.0 program were competent Moodle users, each with over 3 years experience using Moodle in either the role of a student or an editing teacher.

The main priority for the Māori translators of Moodle 2.0 was to ensure an accurate translation of the English source text, reformulated into Māori, but at the same time formulating a comprehensible translation for the average Māori speaking Moodle 2.0 user. There was a certain amount of flexibility required to achieve this balance, as at times staying loyal to the source text to deliver an accurate translation provided an incompatible wordy message that didn't match. For example, the following translation was considered rather wordy:

Source text	*with or without user data*
Translation	*e whai ana, kāore rānei e whai ana i te raraunga kaiwhakamahi*

This was later shortened and changed to:

me ōna raraunga kaiwhakamahi, kore rānei

The translators also had to be wary of translations reformulated to convey a comprehensible message to the user that had changed so much that it was no longer accurate representation of the source text.

Some aspects of the translation work proved very challenging, especially when trying to find equivalent ways of expressing words such as compound verbs. For example:

Source text	*disable-grade-book-history*
Translation	*mono-hītori-puka-māka*

Unlike the English source text, in the Māori translation the sequence of nouns following the verb changes. The real challenge was trying to find equivalent translations to match the source text that sounded natural and fluent in the target language Māori. It wasn't acceptable if the average Māori speaking user of Moodle 2.0 could not discern from either the context or vocabulary what the translation was trying to convey. From experience the translators knew that if the average Māori speaking user of Moodle had to decipher what was being said in Māori, then most would opt to have the interface displayed in English.

Many of the strings required the translation of new words and phrases that had no equivalents in the target language (Māori). Therefore, it is important to develop and exchange common glossaries used by other translators working in this field. The sharing of glossaries, and indeed translations enables translators to keep abreast with the most current and accepted terminology used, however this is a practise not currently undertaken by te reo Māori translators.

Undertaking the Translations

The core translation team consisted of four key members. Hōri Manuririrangi was the project manager and principle translator. Awatea Paterson was a translator. Te Taka Keegan was a translator colleague who provided support on the use of GTT. Tom Roa was a translator colleague who provided translation support. The majority of the translation work was done outside of teaching hours as Hōri Manuririrangi, Te Taka Keegan and Tom Roa are all full-time academic staff. Further assistance was required to meet the 6 month time-frame so Awatea Paterson was brought on as a translator under a fixed term contract. She is an experienced translator that was already familiar with the translation of computer terminology in te reo Māori.

Initially, it was suggested that work could be shared and assigned to other qualified translators who were willing to help, and whilst their contribution would have been greatly appreciated, this offer was later declined in the interests of maintaining consistency with a smaller core translation team.

To be uploaded into the GTT the files had to be converted into a rich text file format. Once uploaded, the files were shared with all four members of the translation team and editing privileges were granted. The only prerequisite being that each member of the translation team had to have their own (free) Gmail account. Once the files had been uploaded and shared on the GTT site, relevant glossaries and shared dictionaries were also uploaded.

The translation work was then undertaken by logging into the GTT, selecting a file to be translated and working through the translations. By opening ‘Show toolkit’ and clicking on ‘Automatic Translation Search’ the GTT automatically performs a search of possible translations of the source text. Translations were generated and displayed for the translator to consider, accompanied with details of the translations origin and a rating to indicate how popular it is.

The bulk of the translation work for Moodle 2.0 began in January 2011 with all of the files completed and verified 6 months later in July. Once the task was completed the translated files were shared with the WCEL team from the GTT site, downloaded and then imported back into the php format that Moodle required. The translations were loaded as the Māori - Waikato language pack of Moodle 2.0.

Feedback From the Translators

On completion of the translation of Moodle 2.0 into te reo Māori the translators involved were asked for their feedback on the use of GTT. Their responses are summarised below.

Learning how to use GTT was relatively easy. It only took a single 40 minute tutorial to show Awatea Paterson the entire process. After two days of translating and moderation tests, Awatea soon became the most efficient user of the GTT in our team. Her progress was noticeable as there is a display of percentage ‘complete’ under each file name.

Perhaps the most relevant feature of the GTT for minority languages such as te reo Māori is the facility to share files, translations, glossaries and dictionaries. Once the resources were loaded into the GTT, sharing them with other translators was simple and the ongoing sharing of translations made the translation task easier.

Another feature of the GTT is its ability to automatically translate uploaded text, either partially or completely, through the use of GTT stored translation memories in Māori. Thus many strings to be translated only required the glancing eye of a translator to verify and amend if necessary. Given the low amount of translated te reo Māori texts the appearance of pre-translated texts were both a surprise and a delight to the reo Māori translators.

The more GTT was used, the more the glossaries were developed, and the translators began to notice that their translations were beginning to appear as automated suggestions from the GTT. The more work that was completed, the more it contributed to enhancing te reo Māori through expanding the shared glossaries and dictionaries that were being used. In turn, the work became easier and the translators became more efficient, thus reducing the amount of time spent having to search terminology. The automated suggestions were non intrusive, and at times gave partial translations or even triggered ideas for possible alternatives. Another added benefit of this automated function was that it constantly provided the translator the opportunity to review ratings of certain translations, which also helped ensure consistency. The Māori language translators stated they felt the Google Translator Toolkit was the most simple to use, effective and convenient translation tool that they had ever encountered.

Summary

This paper has described a case study into how the Google Translator Toolkit was used to assist in the translation of Moodle 2.0 into te reo Māori.

The main convenience of using GTT is the way in which translators can simultaneously edit, save and share the resources to the GTT site. This effectively connects language communities together while negating the need for any storage devices as the material is accessible via any Internet connection. Because the GTT can be accessed by any web browser it means the software is platform independent and can be accessed from anywhere in the world.

Because GTT is available at no cost it is suitable for minority language translators who often operate in environments lacking financial resources. The sharing and real time development of dictionaries, glossaries and translations means that the GTT effectively builds and increases consistencies of minority language resources.

Notes

[1] for a description of issues with the Siraya people see: www.thepetitionsite.com/1/Siraya-reclamation.

[2] see: www.internetworldstats.com/pacific.htm

[3] Te Taura Whiri i te Reo Māori (The Māori Language Commission) – a government commission set up by the Māori Language Act

1987 to promote the use of Māori as a "living language and as an ordinary means of communication".

[4] see: http://translate.google.com/toolkit

References

Austin P.K., Sallabank, J. (Eds.) (2011) *Endangered Languages*. Cambridge: Cambridge University Press.

Keegan, P. J., Keegan, T. T. A. G., & Laws, M. (2011). Online Māori Resources and Māori Initiatives for Teaching and Learning: Current Activities, Successes And Future Directions. *Mai Review*, 1, 1-13.

Lewis, M. P. (Ed.) (2009). *Ethnologue: Languages of the world.* (16th ed). Dallas, Tex.: SIL International. Online version: www.ethnologue.com/. (May 2011).

Nettle, D. Romaine, S. (2000). *Vanishing Voices: the Extintion of the world's languages*. New York: Oxford University Press.

Statistics New Zealand. (2006). *2006 Census Data*. Retrieved from www.stats.govt.nz

Te Puni Kōkiri. (2011). *Te Reo Mauriora Review Report*. Retrieved from www.tpk.govt.nz

Welsh Language Board. (2011). *Advice Note: The Welsh Language, Translation and Technology*. Welsh Language Board, Cardiff.