New Zealand Studies in Applied Linguistics, 2017, 23 (1), 5-16

TUHINGA MĀHORAHORA: TRACKING VOCABULARY USE IN CHILDREN'S WRITING IN MĀORI

Jeanette King, Mary Boyce & Christine Brown

University of Canterbury, University of Canterbury
& Hornby Primary School

Abstract

Māori language and culture immersion programmes have been established now in Aotearoa New Zealand for about 30 years, however there is still not a great deal of research on the proficiency of the children who attend those immersion programmes.

The Tuhinga Māhorahora project has two goals. The first is to test ways of providing timely information to classroom teachers that they can feed back into their curriculum planning and classroom practice. The second is to build a corpus which can provide information of use to those producing curriculum resources in Māori.

The research project is collecting and analysing written texts written in te reo Māori by young learners in Māori immersion settings. The focus is on the vocabulary the learners produce during free writing sessions. These are sessions in which the writers choose their topic and write independently of the teacher. The researchers have collected writing samples into a corpus of approximately 67,200 words to date. We report on our methodology in establishing the database and results and challenges to date.

Introduction

Te reo Māori, the Māori language, a Polynesian language of the South Pacific, is the indigenous language of Aotearoa New Zealand. The language has been the focus of revitalization since the late 1970s when the results of a sociolinguistic survey revealed that very few children were being raised as speakers of the language (Benton, 1991). Initial revitalization initiatives accordingly focused on raising new generations of child speakers. Kōhanga reo ("language nest") Māori immersion preschools were quickly followed by Māori immersion schooling.

Although these schooling initiatives have been operational for over thirty years, apart from work by Cath Rau (2005) and Maraea Hunia (2016), we know very little about the productive language of children in Māori immersion classrooms. However, we do know that there is a significant increase in student reading and writing scores when there is increased support and resources for Māori curriculum development, and teachers' professional development (Rau, 2005).

The Tuhinga Māhorahora project aims to assist teachers in Māori immersion schools by supplying them with information about the words their students are and are not using. The teacher can then lift exposure to underused words and phrases, and introduce alternatives to expand their vocabulary range. Accordingly, as the project continues the aim is to provide evidence based data to support literacy development in Māori immersion settings.

This support is especially important as most teachers are "new" speakers of Māori (Christensen 2003, p. 49), that is, speakers "with little or no home or community exposure to a minority language but who instead acquire it through immersion or bilingual programs, revitalization projects or as adult language learners" (O'Rourke, Pujolar & Ramallo 2015, p. 1). This means that those teaching in the medium of Māori require added support in order to provide a rich linguistic environment for students in their classrooms.

The Tuhinga Māhorahora project is named after the free-writing element in Māori Medium Education (MME) classrooms where teachers are encouraged to allow their students to write for ten minutes a day about any topic they wish (Ministry of Education, 2008). That is, the writing time is not directed by teachers. This writing gives us a window into the child's productive repertoire. While written repertoires are different to spoken repertoires, for logistical and ethical reasons they are much easier to obtain. What children write is typically already within their spoken repertoire so these writing samples provide an insight into both written and spoken proficiency.

In MME settings it is important to ensure quality and quantity of input as part of planning for language success. This is especially important as for many students their only exposure to the Māori language is in school. Rau (2005, pp. 406-407) identifies five groups of children entering MME, ranging from those for whom Māori is their first and only language, through to children who will begin their Māori language learning at school. Despite the fact that most MME schools (including the one in our study) require at least one parent to be a speaker of Māori, Rau found that most new entrants to MME had low levels of Māori language proficiency.

The current project is based on a pilot project implemented by one of the co-authors, Christine Brown, in 2011 and emerged out of her Master of Arts research (2009).

Data

The data in the Tuhinga Māhorahora project currently comprises 1,329 pieces of writing collected from 70 children in year 1-8 MME classrooms during three terms in 2013. In total the database contains 67,168 tokens and 2,100 types. With funding from the New Zealand Institute of Language, Brain and Behaviour (NZILBB), these pieces of writing have been transcribed, tagged and entered into a database. The following is a brief overview of the data collection, transcription and tagging protocols.

At the end of each teaching term the children's writing books were collected and the texts were labeled with participant codes and item numbers and photographed. The photo files were then uploaded to Dropbox. An important feature of this process was that data collection did not disrupt the classroom environment: the writing was produced by the children during regular writing time, and collection occurred out of school hours. Thus the data collection process had a negligible effect on the day-to-day running of the classroom and the school.

The photo files were downloaded by the research assistants and the texts were transcribed and tagged using Xml TEI Editor oXygen (https://www.oxygenxml.com/). A transcription and tagging protocol was prepared and updated as the work progressed. The most current version of the protocol is available at http://www.nzilbb.canterbury.ac.nz/graphics/TuhiMahora-manual-dec15.pdf.

Figure 1 shows a screen shot illustrating the most frequent tag which was used to correct spelling errors, these mainly being incorrect use of macrons. The child's writing appears here in black print with both original and regularized spelling. The tagging appears in blue. As can be seen the original text quickly becomes obscured with the many tags.

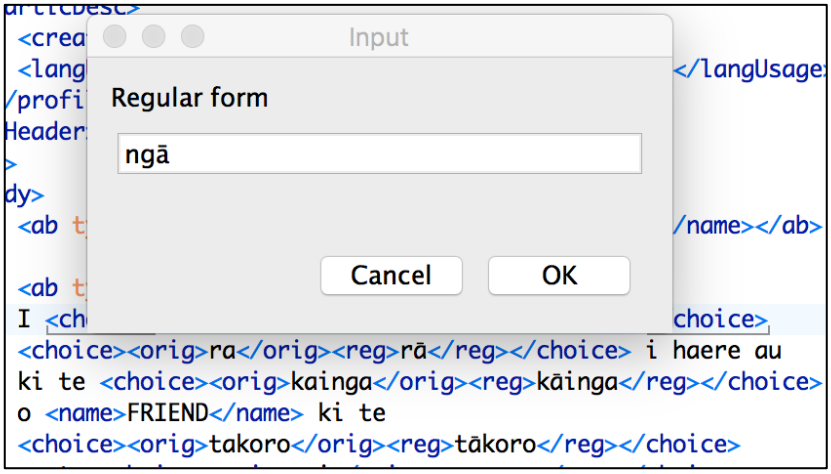

Figure 1. Tagging in oXygen.

Figure 1 includes the dialog box for a plug-in which was produced for the complicated <choice> tagging, used here in the text for the words "rā" and "kāinga", by entering the standard spelling in an entry screen. The <choice> tag was used to retain the child's spelling but allowed for counting forms according to their regular spelling.

It was important to anonymize any personal or place name which could identify the child or school. The bottom of Figure 1 shows how the names of the writer's friends have been replaced with the word FRIEND, as in the tag <name>FRIEND</name>.

The decision to use English words for these replacements ensured that frequency counts for Māori words would not be artificially increased.

Each piece of writing was transcribed into a separate file. We used Text Encoding Initiative (TEI) files as they are a commonly used standard for the encoding of texts in digital form (for more information about TEI see http://www.tei-c.org/index.xml). Information about the file and participant were included in the TEI header via a plug-in, as shown in Figure 2.

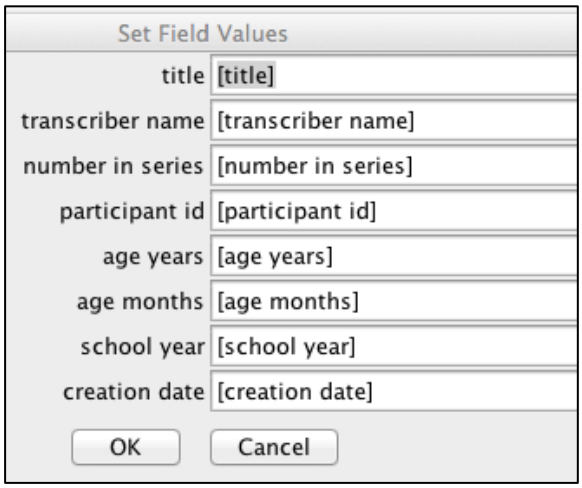

Figure 2. TEI header plug-in.

When completed, the TEI transcripts and photo files of the children's writing were uploaded to LaBB-CAT. LaBB-CAT is a powerful corpus analysis tool developed at NZILBB and originally designed for working with speech files and transcripts (Fromont and Hay, 2012). The Tuhinga Māhorahora corpus is one of the first written corpora to use LaBB-CAT. LaBB-CAT acts as a repository for the corpus and is the platform from which we are able to conduct our analyses. Because LaBB-CAT is an online tool we can work on the corpus on any computer at any time. Access is password protected. The LaBB-CAT software is freely downloadable from http://labbcat.sourceforge.net/.

We can use LaBB-CAT to search for occurrences of words and view them in their context (see figure 3 below). The results of such searches can be exported as a csv file for further analysis. Once uploaded to LaBB-CAT the files (or groups of files) can be downloaded in formats appropriate for use with the WordSmith and Range programs.

In the future further functionality may be added to LaBB-CAT to facilitate additional analyses we may undertake with the Tuhinga Māhorahora corpus.

Analysis

The present analyses use a combination of the various functionalities available in WordSmith, Range and LaBB-CAT. WordSmith enables us to produce frequency

lists, allowing us to identify words for further analysis. Range, developed by Alex Heatley and others, enables us to compare word usage by the children against frequency lists compiled by Brown (2009).

For this paper we have selected data from the Year 3 classroom as an example of the types of analyses which can be performed and how this information can be used to assist teachers. This year group has been chosen for two reasons. Firstly, the children have passed through the emergent writing stage and some are writing more extended pieces. Secondly this is the group for which we have the most data: 365 pieces of writing produced by twelve children aged from 6 years 8 months to 8 years and 1 month old at the time of writing. The texts ranged from 5 to 189 words, with an average length of 41 words.

Range

Range is a computer program designed to analyze the vocabulary load of texts according to frequency bands (Heatley et al., 2002). This is achieved by the use of frequency lists which can be formulated by the user (Range comes with English frequency wordlists). Range can compare vocabulary use in up to 32 different texts at a time against the frequency lists.

For our analysis we compared the children's use of words in relation to eleven wordlists compiled by Brown (2009) which contained the most frequent words in Māori. Nine of the wordlists consist of content words, (1820 words in total), ranging from the most frequent (list one) to the least frequent of these words (list 9). In addition, there is one list containing function words (157 words) and one containing a list of names the children are commonly using. Range also collates the words used by the children which are not in any of the lists. This enables us to easily see the English words the children are using, showing the Māori vocabulary that the children need.

Wordlists 1 to 9 were constructed using several corpora of adult language use, totaling nearly two million words, including readers written for children in MME environments. These lists were then moderated for classroom language use by Christine Brown in consultation with teachers who identified common words in use in the classroom context.

Table 1 shows the results from Range for the Year 3 children. The largest proportion of words used by the children are function words (58% of the texts), a proportion which is roughly consistent with other Māori texts such as the Māori Broadcast Corpus (65% function words) (Boyce, 2006) and the texts used for Brown's analysis (62% function words) (Brown, 2006). Māori, as with most Polynesian languages, uses a large range of function words to indicate the various grammatical roles of content words (plurality, tense, etc.) (Harlow, 2006:24).

Table 1 Range results for Year 3 students

word list	tokens	tokens as % of text	types
one	2980	20.86	114
two	421	2.95	87
three	360	2.52	62
four	350	2.45	75
five	140	0.98	40
six	66	0.46	23
seven	351	2.46	33
eight	55	0.39	27
nine	26	0.18	12
function words	8,264	57.85	99
names	539	3.77	58
not on lists	733	5.13	316
Total	**14,285**	**100**	**946**

The second highest number of tokens (21%) is found in word list one which contains the 134 most frequent words.

The higher than expected use of words in list seven is because this list contains the words for the months of the year and most of the Year 3 children begin each piece of writing with a formulaic date phrase which includes the month.

Table 1 also shows the number of types used from each word list so we can calculate what proportion of words on each list the Year 3 children are using. In this case they are using 85% of the words in list one, but only 43% of function words.

WordSmith

Using WordSmith (Scott, 2004), we can also look at overall word frequencies amongst the writing of the children.

Table 2 Raw frequency list for Year 3 students

Number	Word	Frequency	%	Texts
1	TE	1559	10.2	12
2	I	1430	9.4	12
3	KI	689	4.5	12
4	ME	651	4.3	11
5	O	428	2.8	12
6	HAERE	421	2.8	12
7	RĀ	400	2.6	11
8	KA	366	2.4	12
9	KO	355	2.3	12
10	NGĀ	338	2.2	12
11	A	319	2.1	12
12	AU	229	1.5	12
13	AHAU	219	1.4	11
14	HE	179	1.2	11
15	PAI	150	1.0	11

As shown in Table 2, the top 15 words used by these twelve year 3 children are mostly function words, with only three content words (shaded).
The frequency column shows how many tokens of each word occurred in the children's writing. The far right column shows how many of the twelve children used each word. We can see that the top 15 words were produced by nearly all twelve children.

If we remove function words from the list (along with names of the months), we can see the 15 most frequent content words (Table 3). Again, these words appear in the writing of almost all of the children.

Table 3 Most frequent content words for Year 3 students

Number	Word	Frequency	%	Texts
1	HAERE	421	2.8	12
2	RĀ	400	2.6	11
3	PAI	150	1.0	11
4	TĀKARO	121	0.8	12
5	KAI	119	0.8	11
6	WHARE	107	0.7	11
7	WĀ	89	0.6	12
8	WHAKATĀ	76	0.5	11
9	RUNGA	72	0.5	11
10	MAHI	70	0.5	11
11	MURI	69	0.5	10
12	MEA	66	0.4	10
13	ROTO	66	0.4	11
14	TIKI	66	0.4	11
15	WHAI	47	0.3	6

All but two of these words appear in frequency list one, the most frequent words (Brown, 2009).

Feedback to teachers

The information obtained from these analyses can be used to provide insight for the teachers.

Looking at Table 3 one item that stood out for further analysis to those with a knowledge of Māori is the eleventh most frequent content word “muri”. “Muri” is a location word referring to “behind” (when talking about space), but meaning “after” (when talking about time). We can see from the frequency column that there were 69 instances in this corpus of year 3 writing, and the right-hand column shows that ten out of twelve children in the class were using this word.

We can use LaBB-CAT to look at how these students are using “muri”. Table 4 only shows ten of the instances (one from each child who uses the word), but they are indicative of all 69 instances. Note that while spelling mistakes have been corrected, grammatical errors have not.

Table 4 Uses of "muri" by the Year 3 students.

Student code	Example
124	**Whai muri** i te kura kei te haere au ki te whare
125	Ka tiki te hōanga, **whai muri**, ka peita
126	I haere mātou ki te kai sushi. **I muri i tērā** i haere ki te warewhare
127	I moe ki tōna whare me **whai muri i tērā** ka kai
128	He pai tērā ki ahau. **Ā muri i tērā** i haere mātou
130	I kai ahau ngā rare maha. **Ā muri** te kura ka haere
131	Ka tākaro ki waho **whai muri** i te tīni kākahu
132	I tatari a Obi-wan Kenobi me a Qui-gon me a Darth Maul. **Whai muri tērā** i tapahi a Darth Maul i a Qui-gon Jinn
133	Ka haere ki te whare karakia. **I muri i tērā** i te wā i tiki aihikirīmi
134	**Whai muri** i te kura ka haere au ki te kauhoe

All of these examples refer to time, mostly in the phrases "whai muri" – equivalent to "following on", and "i muri i tērā" – equivalent to "after that". In other words, "muri" is being used exclusively for time and sequence cohesion. The use of these phrases is a good example of how formulaic expressions can be useful building blocks in language expression (see Wray's "needs only analysis" 2002, and King, 2015).

This is an example of how there can be discussions with the teacher about how to model a wider range of cohesive devices.

Besides looking at words or phrases that the children are using, we can also look at words in the top frequency lists which are not being used by the children in their writing. Knowing which high frequency words are not being used assists the teacher to plan to lift learners' exposure to these items.

As well as content words we can look at strengthening the students' use of function words. For example, "kāore" is a word used to negate verbal and location sentences in Māori. In the year 3 texts there were only six instances produced by three of the twelve children. Four of these do not use the "kāore" construction accurately. Perhaps this is developmental, but it could be lack of exposure to correct forms. A discussion might lead the teacher to consider whether and how to address this in the classroom.

Reflection

In the pilot conducted by Christine Brown in 2011 she transcribed the children's writing into a running text file at the end of each teaching term. She was able to use WordSmith and Range to provide timely feedback to teachers. In addition to feedback as per the types of analyses above, she was also able to give the teachers other useful information. For example, comparisons with the word lists indicated the high frequency words that children did not use. That information was used to encourage vocabulary growth in these "high value" words. In addition, words which the children used but were not on the high frequency lists were good indicators of children's interests, activities and experiences out of school. This is valuable information for teachers to connect with children's lives. Dialect preferences also become evident, and were able to be supported.

Teachers studied the English words used by the children when they didn't have a Māori word in their vocabulary. They were then able to incorporate the Māori equivalents as target words into shared writing sessions. This resulted in reference pieces of writing which were displayed on the classroom wall. These pieces were then referred to often throughout the year. Grammatical errors were also analyzed and resources were made to support correct use in both written and spoken activities.

The pilot encouraged teachers to think more specifically about the words their students were using and those that needed to be developed. Analyzing children's writing in this way provided a rich and diverse fresh evidence base which provides good direction and motivation for focused teaching.

When compiling the Tuhinga Māhorahora corpus in 2013 we severely underestimated the time it would take to transcribe and tag the texts. Accordingly we were unable to provide feedback to the teachers in a timely manner as in the 2011 pilot.

We are currently examining ways in which we could make the feedback more effective. One way would be to substantially reduce the amount of tagging. In particular we could regularize the children's spelling during the transcription process without retaining the original spelling since the analysis of spelling mistakes is not a primary aim of the project. This would greatly simplify the transcription and tagging process. There are pros and cons for all transcription and tagging decisions and while standardizing spelling during the transcription process would be quicker in the short term it is less flexible for later purposes.

In addition, in many classrooms students are now writing directly on tablets. Capturing digital data would also greatly expedite the formatting of text in preparation for analysis and allow us to deliver information to teachers more efficiently during the school year. We are also keen to identify other computational methods or tools that might assist.

Now that we have tested and adapted our protocols we intend to apply for funding from the Ministry of Education to enable us to achieve our aim of providing evidence based language support for teachers and students in Māori immersion classrooms. In this way we will be able to increase the database to a size where it can form a useful reference point to ensure curriculum materials are developed at the appropriate levels for students in Māori immersion schooling. A large corpus of children's productive language would be an excellent resource for language planning and curriculum development for this endangered language.

At present there is no national database of children's productive language in Māori and very little is known yet about the stages of language development for children in Māori medium settings.

Acknowledgements

We would like to acknowledge the children, their families, teachers and school who agreed to have the children's writing collected. We are very grateful for financial support from the New Zealand Institute of Language Brain and Behaviour (NZILBB) as well as the support of NZILBB's Software Programmer Robert Fromont, and Research Technician, Scott Lloyd. The project could not have proceeded without the work of research assistants Roberta Tainui, Caitlin Swan and Niwa Wehi who transcribed and tagged the written material.

References

Benton, R. A. (1991). *The Māori language: Dying or reviving?* Honolulu: East West Center. Reprinted by New Zealand Council for Educational Research, 1997.

Boyce, M. T. (2006). *A corpus of modern spoken Māori.* Unpublished doctoral dissertation. Victoria University of Wellington, New Zealand.

Brown, C. M. (2009). *Assessing the Readability of Māori Language Texts for Classroom Use.* Unpublished Master's thesis. University of Canterbury, New Zealand.

Christensen, I. S. (2003). Proficiency, use and transmission: Māori language revitalisation. *New Zealand Studies in Applied Linguistics,* 9(1), 41-61.

Fromont, R., & Hay, J. (2012). LaBB-CAT: an annotation store. *Proceedings of the Australasian Language Technology Association Workshop*, 113−117. Retrieved from http://www.aclweb.org/anthology/U/U12/U12-2015

Harlow, R. (2006). *Māori: a linguistic introduction.* Cambridge: Cambridge University Press.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/about/staff/paul-nation - vocab-programs

Hunia, T. M. (2016). *He kōpara e kō nei i te ata: Māori language socialization and acquisition by two bilingual children: a case study approach*. Unpublished doctoral dissertation. Victoria University of Wellington, New Zealand.

King J. M. (2015). Metaphors we die by: change and vitality in Māori. In E. Piirainen & A. Sherris (Eds), *Language Endangerment: disappearing metaphors and shifting conceptualizations* (pp. 15-36). Amsterdam: John Benjamins.

Ministry of Education. (2008). *Te Hōtaka Tuhituhi Māhorahora.* Retrieved from http://eng.mataurangamaori.tki.org.nz/Media/Files/Manu-tuhituhi/Te-Hotaka-Tuhituhi-Mahorahora

O'Rourke, B., Pujolar, J., & Ramallo, F. (2015). New speakers of minority languages: the challenging opportunity - foreword. *International Journal of the Sociology of Language, 231*, 1-20.

Rau, C. (2005). Literacy acquisition, assessment and achievement of year two students in total immersion in Māori programmes. *International Journal of Bilingual Education and Bilingualism, 8*(5), 404-432.

Scott, M. (2004). *Oxford WordSmith tools version 4.0*. [Computer Program, manual and associated files]. Oxford: Oxford University Press. Retrievable from http://lexically.net/wordsmith/

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.