Tuhinga Māhorahora Project
Style Guide, Transcription and Mark-up Manual

Brief Description

This document records and presents the methods used to collect and process data in the Tuhinga Māhorahora project.

The Tuhinga Māhorahora

The Tuhinga Māhorahora Project is collecting free and dictated writing in Māori from writers in Māori-medium classrooms. The initial phase is collecting from years 0-8 in selected schools and from a single city as a pilot for a more extensive project. It is hoped that the project will span all years of the compulsory education sector and be nationwide in New Zealand. This, of course, is contingent on funding.

This document is designed to assist the team transcribing, adding markup, editing and proofreading the transcripts. It presents a summary of the practical aspects of the methods and processes of data collection, processing and management.

This document provides details of the required mark-up and style choices for the transcription of texts for the Tuhinga Māhorahora corpus, and their entry into the project database.

The mark-up protocols are based on the tag set developed for the Lancaster-Leverhulme Corpus of Children's Writing (LCCW). See Smith, McEnery and Ivanic 1998 'Issues in Transcribing a Corpus of Children's Handwritten Projects' *Literary and Linguistic Computing* Vol. 13 No. 4 pp. 217-225 for details. It is also based closely on *Barebones TEI - A Very Very Small Subset of the TEI Encoding Scheme*, C. M. Sperberg-McQueen. Document No. TEI U6. 30 Aug 1994. rev. June 1995. Additional TEI items are added as required.

This document (electronic) is dynamic - it will evolve as the project proceeds and be amended regularly as required. The TM Project Manual manager is responsible for writing and maintaining the manual. All queries and suggestions for changes are sent to the TM Project Manual manager, as are any items to be included in the appended style guide.

Principles for transcription

Overarching principle

Our aim is to accurately transcribe each individual text and add it to the Tuhinga Māhorahora corpus.

We will stay true to the writer's output while also ensuring consistency across the texts transcribed for the corpus to facilitate analysis.

We will add mark-up to each text to facilitate later analysis of that text, and to provide metadata about each individual text. The general metadata for each text will be encoded in the TEI header for that text.

Some more specific guidelines

1. Keep the mark-up as light as possible, only tagging features which are important for our analysis. Nevertheless, the aim is to have a mark-up strategy for all aspects of the written texts that may be of interest, but within the constraint of keeping it light. This means a bigger set is recorded in this manual than is likely to be needed for every text.
2. Some features will only probably occur in different 'ages/stages' of writing development. For example, the example below about the <note> tag is likely to only occur in the writing of very young children.

   Illustrations may occur at any age, but the nature of these will change with degrees of sophistication. For example, emerging writers will often express meaning through their illustrations and speak that meaning to their teacher, who records it in writing which the child then copies over, or copies independently. A more sophisticated writer may include an illustration / graph / table in addition to written text. The LCCW mark-up tags for illustrations are used.
3. Early instances of transcribing will likely yield features as yet not detailed in the guidelines. These are to be raised and a decision made as soon as possible and added to this manual by the TM Project Manual manager.
4. Transcriber and those doing the mark-up will follow the mark-up style guide, and query anything with the TM Project Manual manager (currently Jeanette King) as soon as an issue arises. This is necessary so that we do not have too much repeat mark-up checking to do.

   The TM Project Manual manager will make changes to the style guide if / when necessary. For example:

   - Transcribers/annotators notice a feature in the writing not covered by the style guide.
   - Transcribers/annotators notice that a mark-up strategy is too blunt / not fine-tuned enough to capture enough detail.
5. The phases in text preparation are as follows:
   - the original texts are coded with writer identity and item number and photographed using the project's iPad mini. These jpg files are uploaded to the TM Project dropbox Transcripts folder.
   - The jpg photo files are named with the identity code of the writer, plus the code for which piece of writing this is from that writer, e.g. in the filename 09324-017, the 09324 would be the writer identity code, and the 017 after the hyphene would indicate that this is the 17th text from that writer.
   - the photographed text is transcribed and marked up[1] (formerly by Roberta Tainui, now Niwa Wehi and Caitlin Swan) using the TEI Editor oXygen and the Tuhinga Māhorahora xml template file. TEI header information is also completed. The resulting xml file is saved with its original name plus 'mrkp',

[1] In the initial stages of the project the photo files were then transcribed using text edit and saved as a UTF-8 TM Project dropbox folder by Mike Davey.

e.g. 09324-017-mrkp.xml. This file is uploaded to the TM Project dropbox, into the Markup folder, as appropriate.

- the marked-up text is now checked for accuracy of markup by a third team member (currently Jeanette King). The finalised files and accompanying photo files are saved onto the NZILBB server and uploaded into LaBB-CAT, along with demographic information of the participants.

Style / template for transcripts

Transcription and annotations are made using the oXygen XML Editor, and using the Tuhinga Māhorahora template.

Before using oXygen a plugin to help deal with the regularisation tags needs to be installed:

Plug in for regularisation tags

1. Go to this website: https://labbcat.canterbury.ac.nz/download/ and download the file called *teiregularization.jar*
2. Find the folder where Oxygen is installed on your computer - if it's Windows it might be something like *C:\Program Files\Oxygen XML Editor*
3. In that folder, there's a subfolder called *plugins*
4. In the *plugins* subfolder, create another folder called (something like) "teiregularization"
5. Copy the *teiregularization.jar* file you downloaded into the new folder you just created.
6. Double-click on it. This should open a window that has a message saying "File extraction complete" and in the folder you created, there should be a new file called *plugin.xml* next to the *teiregularization.jar* file.

Working with the TEI Editor

For each new transcription/annotation, open the Tuhinga Māhorahora xml template.

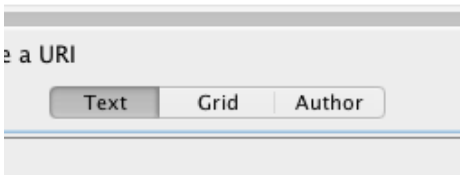
1. There are three editing modes:

"Text", which shows all the XML tags explicitly,

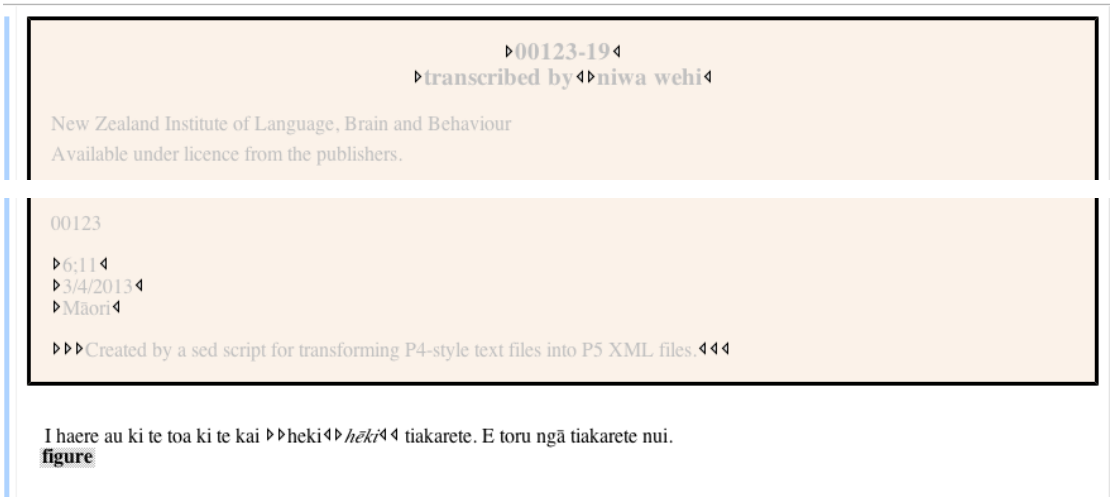
"Grid" (ignore, as it's not useful for us), and

"Author", which selectively hides the tags and displays text with colour/font styles to make it more readable.

2. You can change between views using the selector below the text editor:



3. The author view looks like this:



4. The blank document isn't totally blank - it has "Text here" as a place-holder for your text. So the first step is to replace that with the actual transcript.

5. Go to the "Text" view, and select "Text here".

6. Copy and paste in the text of the transcript file, if this already exists in another format. Otherwise enter your transcription here. Note that the transcription does not attempt to mimic the line breaks in the child's text. For example:

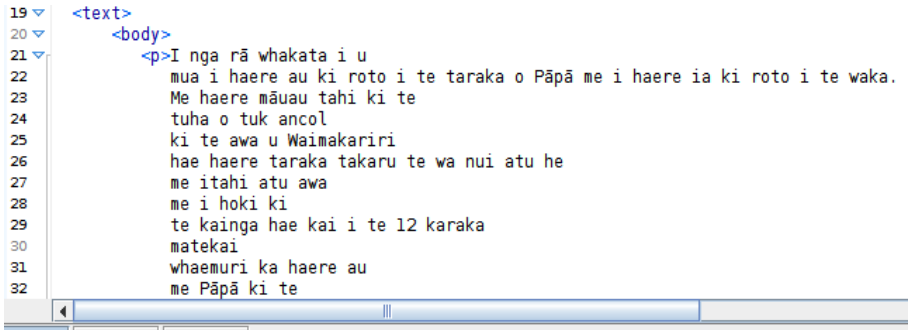
Child's text:

I haere au
ki te toa ki te
kai hēki tiakarete.

In xml:

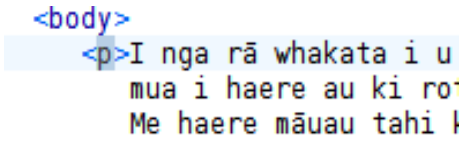
<ab type="free writing"> I haere au ki te toa ki te kai hēki tiakarete.</ab>

6. Transcription and annotation will proceed as below. TEI header information will be added, see Appendix four.



7. If the text is wrapped in a <p> tag - change it to an "anonymous block" <ab> instead:

Select the "p" inside the tag.



Type "ab" instead. Use the following type attributes: free writing, dictated writing, or directed writing in each <ab> tag.

Free writing is writing that the child has produced by themselves (as far as we can tell). This is the majority of the writing.

<ab type="free writing">I te <sic>aniw</sic> ki au haere </ab>

Dictated writing usually appears only in the year 1 writing. It's where the child appears to have dictated all or part of their story to the teacher who writes it in the book for the child to copy underneath.

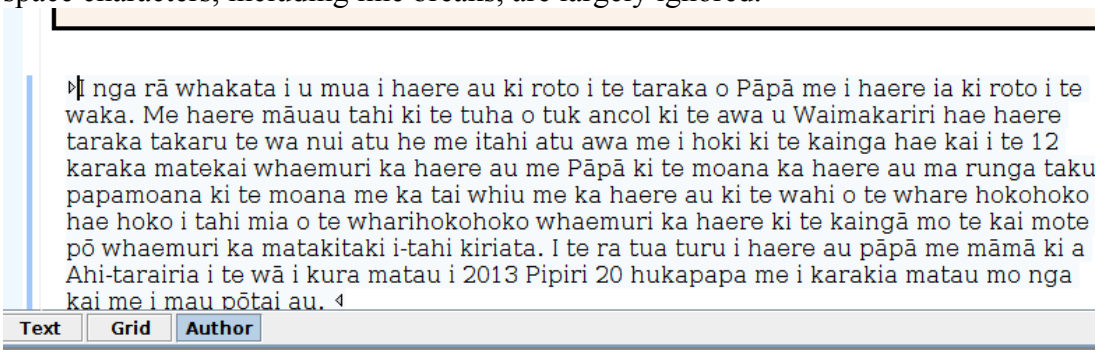
<ab type="dictated writing">I te wā whakatā i haere au ki Makitanara</ab>

Directed writing is where there is **clear** evidence that the text has been influenced by the teacher. For example, when all the children in class on a particular day follow a set topic format.

<ab type="free writing">Ko te kai pai ki ahau ko te panikeke</ab>

8. If you scroll to the bottom, you'll see that the end tag has automatically been changed to </ab>.

9. Note that any date, used as heading to the main body of the text will also be surrounded by its own separate <ab> tag.
10. If you go to "Author" view, you'll see that all the text appears in one big mass, instead of divided up into the lines you typed. This because, with XML, white-space characters, including line breaks, are largely ignored.



Annotation

Now you can annotate the text. This process is as easy to do, and works the same, in "Text" or "Author" view.

Some general guidelines for tags

1. See the explanations below for how to use each tag, and also see the summary table the Tuhinga Māhorahora Project tagset in Appendix 5.
2. Every tag must be accompanied by a closing tag which includes a forward e.g. <note> … </note>. Note that Oxygen automatically inserts the closing tag for you.
3. Tags are nested when necessary. More specific tags are nested within broader tags, e.g. a <foreign> … </foreign> tag may occur around a specific word within a <name> … </name> tag.
4. Punctuation is included only if it is present in the writing.

The tag descriptions below are presented in order from most to least frequently used.

TEI Header <teiHeader> … </teiHeader>

Each transcript will have a TEI header that stores the metadata for that file (e.g. writer code, school code and so on). See Appendix 4 for the template. The purpose of the TEI header is to provide information for users of the corpus and to enable analysis of groups of written material which can be grouped by any one of the codings in the header.

Anonymous block tag <ab> … </ab>

This tag marks every new block of text. We are using this following the LLCW system rather than the typical <p> paragraph tag used in TEI. It does not presume that a chunk of text is a paragraph. It covers headers and the like too, so is a blunter tag.

In practice we use this tag to separate out the date - which the child usually writes at the top of most pieces of writing, and also any obviously deliberate paragraphing of text – most likely amongst older children.

<ab type="free writing">13 o Mahuru</ab>

Use the following type attributes: free writing and dictated writing in each <ab> tag.

> Free writing is writing that the child has produced by themselves (as far as we can tell). This is the majority of the writing.
>
> Dictated writing usually appears only in the year 1 writing. It's where the child appears to have dictated all or part of their story to the teacher who writes it in the book for the child to copy underneath.

<ab type="free writing">I te <sic>aniw</sic> ki au haere </ab>

<ab type="dictated writing">I te wā whakatā i haere au ki Makitanara</ab>

If the whole text, apart from the date line, is something the child has copied (e.g., a waiata) then that file is removed and put into a 'deleted file' folder with information and the reason why the file has been removed put onto the deleted file excel sheet.

Regularisation tag

We will correct spelling of Māori words that do not conform to the *Te Taura Whiri i te Reo Māori Orthographic Conventions* by using a regularisation tag. The Te Taura Whiri Conventions are available online at:

http://www.tetaurawhiri.govt.nz/english/pub_e/conventions.shtml.

In practice, the main use of this tag is to correct macron use. Use the online Māori Dictionary at http://www.maoridictionary.co.nz/ to check macronisation and the spelling of words. The only exception to this is the writing of the names of the months of the year. In all cases these will be spelt as one word without hyphens.

1. Make sure you're in "Text" mode (not "Author" mode).
2. Select a word that needs regularization.
3. Hit <Ctrl>+<R> on your keyboard. You should be asked for the regular form (the default is a copy of the original form).
4. Enter the regular/correct form, and click OK or hit <Enter>
5. In the XML, it should have inserted something like:
   <choice><orig>matua</orig><reg>mātua</reg></choice>

Another use of this tag is to break up words that the child writes as one word or to put together words which the child writes as separate words.

> he <choice><orig>rekarawa</orig><reg>reka rawa</reg></choice> atu
> haere mai <choice><orig>ki a</orig><reg>kia</reg></choice> kite i ahau

We will also use regularisation tags to correct the spelling of any English words or names.
I kai te <foreign xml:lang="en"><choice><orig>puding</orig><reg>pudding</reg></choice></foreign>

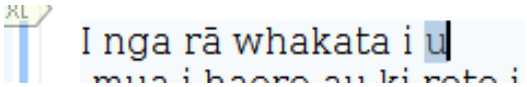
English words such as 'mmmm' will be regularised to be spelt 'mmm'.

Tags wrapped around text

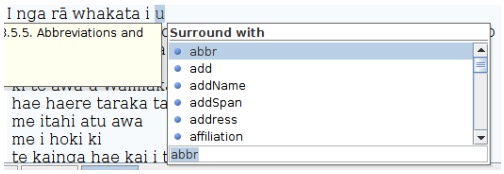
The general principle for working with tags which are wrapped around text is to use a

<Ctrl>+<E> keystroke (⌘ + E on a Mac). The same option is available in the menu: *Document|Markup|Surround with Tags*). The main tags which are wrapped around text are <name> <unclear> <sic> and <quote>.

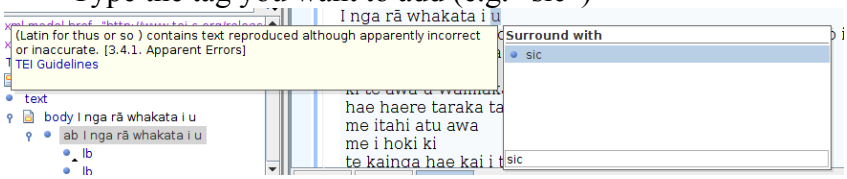
1. Select a word you want to wrap a tag around (e.g. the "u" at the end of the first line)



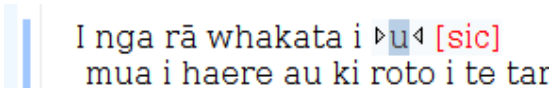
2. Type <Ctrl>+<E> (or ⌘ + E on a Mac) on your keyboard. A list of possible tags will appear.



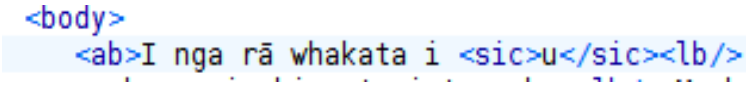
3. Type the tag you want to add (e.g. "sic")
4.



5. Hit <Enter>. The opening and closing tags have now been added. You can see it in "Author" view:



...and in "Text" view:



Name tag <name> … </name>

The basic use of this tag is to enable us to separate out names from any vocabulary analysis. We define names as being:

Personal names, including Māmā and Pāpā when used as names (see below)

Place names

Names of days of the week and months of the year

Movies, books, computer games, song names

An example of the two ways the words māmā and pāpā can be used as either names or descriptors in noun phrases are:

Not names (because there is a definitive, eg, te, taku, tana, etc before the word māmā or pāpā. This is equivalent to the English usage 'my mother':

i haere taku māmā me taku pāpā me a <name>FRIEND</name>

Names, because the particle ‘a’ or ‘ko’ should be in front. This is equivalent to the English usage ‘Mum’.

> he pai ki ahau rāua ko <name>pāpā</name> me <name>māmā</name>

Or when the parents are being directly addressed:

> i te pātai ahau ki aku mātua "<name>Māmā</name> <name>Pāpā</name> ka taea e <name>FRIEND</name> te moe mō te pō?

An example of the name tag for a place name is:

> <name>Te Whanganui-a-Tara</name>

Note that some place names will also need regularisation tags to indicate any spellings that do not conform to the *Te Taura Whiri i te Reo Māori Orthographic Conventions*.

The regularisation tags will be nested within the <name> tags.

The <name> tag can be tailored, as shown below, to anonymise potentially revealing information, such as (but not limited to) the name/s of the writer, members of their family, friends and school.

> Ko<name>FRIEND</name>taku hoa.
>
> I haere mātou ki<name>SWIMMINGPOOL</name>ki te kaukau.

Note that in the second example where the name of the pool has been replaced, this is because the use of this name would indicate that the child attends a particular school which is located adjacent to this facility. Note that the replacement word indicating the nature of the name which has been removed is written in capitals and in English. This is to ensure that these words are easily removed from word lists produced in the analysis phase. The current list of replacement words are:

<u>People</u>

ADULT (for adult who isn’t otherwise included in the list below)

AUNT

AUTHOR

BROTHER

COUSIN

FATHER

FRIEND

FRIEND’SPARENT

FRIEND’SSIBLING

GRANDFATHER

GRANDMOTHER

MOTHER

NAME (for any other person’s name, who isn’t otherwise included in this list)

PET

RELATIVE

SIBLING

SISTER

STUDENT (i.e., others in class)

SURNAME

TEACHER

UNCLE

UNKNOWN

Places

AMENITY (that doesn't fit into any of the other categories here)

CLASS

CHURCH

HOTEL (when a particular hotel is associated with a family member, eg they own it)

OVERSEASLOCATION (that a child has visited, not usually for places in Australia, but further afield where the fact that they have visited there could be identifying)

PARK (when child describes a visit to a park during school)

PRESCHOOL

PRIMARYSCHOOL

SCHOOL

SCHOOLBUILDING

SHOP

SHOPPINGCENTRE

SUBURB

SWIMMINGPOOL

TEAM (for names of sports teams or kapa haka groups)

A reminder that it won't always be necessary to anonymise some of the names in these categories if they won't potentially reveal the child's identity. In this example the name of the shopping mall has been retained.

I haere mātou ki <name>Te Whare Pukapuka o Papanui</name>

Movie names and names of fast food outlets are also names:

I haere mātou ki <name>Makitānara</name>

But McDonalds, for example, is not a name when it's referring to the food itself:

I kai makitānara mātou

Morpheme tag <m> … </m>

If a child uses a word which is a blend of Māori and English elements use the following coding to tag the different language elements of the word. In this case the child has put the comparative ending ‘er’ on the Māori word ‘tere’ > ‘tereer’. The following coding allows the base word to be counted in the frequency list for ‘tere’.

<w><m xml:lang="mi" baseForm="tere">tere</m><m xml:lang="en">er</m></w>

- the <w>...</w> marks out the whole word.
- each <m>...</m> marks out a morpheme - these end up on the new "m" layer in LaBB-CAT, and also they're concatenated together to make up the word on the "transcript" layer.
- the xml:lang attribute specifies the language of the morpheme, and this ends up being marked in the "language" layer (this currently looks a little wonky on the transcript page, but it's represented correctly in the database, so future analysis, if desired, should be doable).

The baseForm attribute specifies the form of the word that should appear in your statistics (this will end up on the "orthography" layer from which the stats are calculated). This is the same as the morpheme text in this case, but it could conceivably be different.

Unclear tag <unclear> … </unclear>

The <unclear> tag is used to mark up any material that is too unclear to be transcribed accurately. Information about why the text is unclear can be added, if necessary, by using some of the attributes associated with the <unclear> tag. For example, there is an attribute tag for expressing how confident you are about the transcription:

te whare <unclear cert=“medium”>o taku tupuna</unclear> nā taku matua i hanga

And also the reason for the lack of clarity:

te whare <unclear reason=“very faint writing”> o taku tupuna</unclear> nā taku matua i hanga

Note that you can use more than one of these attribute tags, if required.

Note that there doesn’t need to be any text between the opening and closing <unclear> tags.

te whare <unclear reason=“illegible”></unclear>nā taku matua i hanga

We will also follow the LCCW protocol of giving the writer the benefit of the doubt and not use explicit tags where items are only partially unclear and can be recovered with reasonably high confidence (Smith, McEnery and Ivanic, 1998 p. 222).

Sic tag <sic> … </sic>

This tag is used to indicate exact wording where it is left as in the original but is not a recognisable spelling error. Smith, McEnery and Ivanic, 1998 p. 222 give an example where a word is repeated. We have had an example where a child used an unknown word ‘mahangata’ where it was unclear what word or words were meant.

i ētahi <sic>mahangata</sic>te mutunga

Quote tag <quote>

This tag is used when the writer is quoting the well known words of another writer or speaker. A good example would be when quoting a well-known proverb. This allows words within the quote tag to be excluded from analysis.

> Nō reira <quote>me whai i te iti kahurangi</quote>mēnā ka pīrangi koe …

Note that this tag is **not** used for reported speech. Just transcribe any punctuation which the child may, or may not, use.

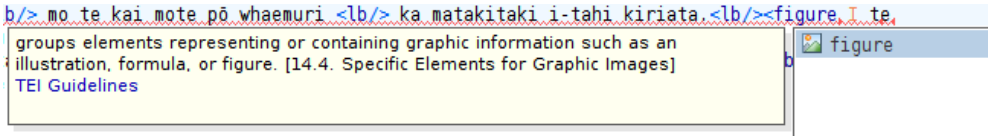
For example

> I kī mai taku māmā, "Haere atu."

Tags not wrapped around text

The easiest way to add tags that aren't wrapped around text (e.g. <figure> and <note>) is in "Text" view.

1. Go to "Text" view
2. Place the cursor at the point in the text where there's a picture.
3. Type "<" with the keyboard. You will see a list of all possible tags.
4. Type "figure".



5. Hit <Enter> and the tag will be added.

Note tag <note> … </note>

This tag is to mark up any comments or notes added by team members handling the transcripts. Here is an example of one that might be made by a transcriber:

i haere <note>the word "au" has been added by the teacher</note> ki a Tāmaki-makau-rau

When uploaded into our software the program will recognise text between note tags but will not include it in any analysis.

Another use of this tag is to include onomatopoeic type words written by the child:

Ka umere <note>the child writes the following onomatopoeic words: a a a a a a</note>tētahi āwhina

Figure tag <figure></figure>

This tag marks any graphic element produced by the writer or any textual or visual material imported into the text from an external source, e.g. a magazine cut out or a photograph. (See example on p. 220 of Smith, McEnery and Ivanic.) A note may be added to give more information about the nature of the graphic, and marked up with the <note> tag, though in general we will just insert the <figure> tag by itself.

Except when there is text in the figure (see below), it makes no difference whether the figure tag is included within an <ab> tag or not.

So, two uses. For a drawing the child has done:

<figure></figure>

For a photograph or magazine picture which has been pasted in:

<figure type="photograph"></figure>

If there is text written in the drawing (other than just names of people), make sure there is an <ab> tag around the figure tag (that is, it is a separate anonymous block) then just write the words between the figure tags.

<ab type="free writing"><figure>kei te kai a māmā </figure></ab>

If the child has people speaking in the picture use an <ab> tag nested outside the figure tag with a quoted text <q> tag nested inside that.

<ab type="free writing"><figure><q>Titiro ki tēnei!</q></figure></ab>

If the speech is in a speech balloon .

<ab type="free writing"><figure><q type="balloon">Titiro ki tēnei!</q></figure></ab>

If there is more than one speech balloon, include them all within the <ab> tags:

<ab><figure><q type="balloon">Titiro ki tēnei!</q><q type="balloon">Anei ahau</q></figure></ab>

In general, we are not writing descriptions of the figures, but if anything needs to be mentioned it can be done using a <note> tag (see below).

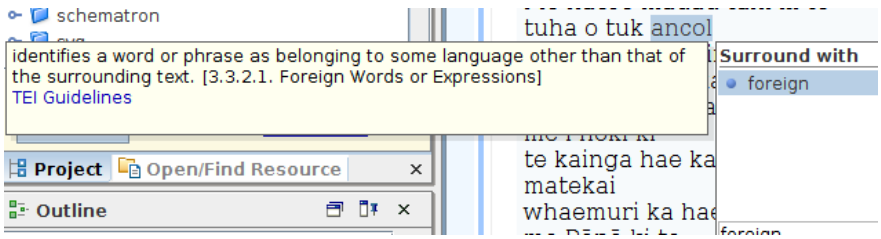
Tags with attributes

For tags that have attributes, like <foreign> for marking English (or other non Māori) text. This tag will often be nested within another tag, for example, <name>.

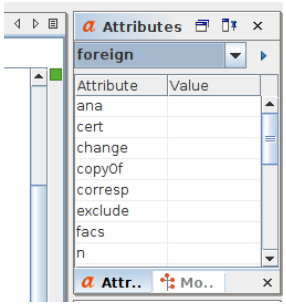
Language tag <foreign>

This tag is used to mark up any content that is in a language other than Māori. See Appendix 3 for the language codes.

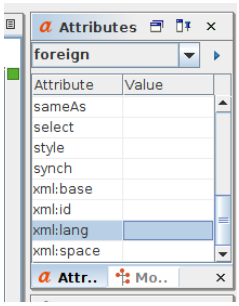
1. Tag the word as above, using <Ctrl>+<E> (or ⌘ + E on a Mac) and entering "foreign".



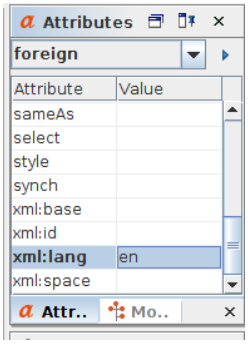
On the right hand side, there's a panel called "Attributes", which displays all the possible attributes of the current tag.



2. Scroll down until you see the attribute you need (in this case “xml:lang”).



3. Double click in the value column and enter the value of the attribute. For *xml:lang*, type the two-letter ISO code “en”



The result will be something like:

<foreign xml:lang=“en”>my pet</foreign>

<foreign xml:lang=“sm”>fa’afetai lava</foreign>

The following are two nested examples where names, or parts of names are in English:

<name><foreign xml:lang=“en”>Justin Beiber</foreign><name>

<name>Whirimako<foreign xml:lang=“en”> Black</foreign><name>

Punctuation tag <pc>

This tag is used to mark up any emoticons written with punctuation characters or kisses, etc. written as xxxx.

1. Tag the word using <Ctrl>+<E> (or ⌘ + R on a Mac) and entering “pc”. This puts the <pc> tag around the symbols.
2. On the right hand side, there’s a panel called “Attributes”. Scroll down until

you see type. Double click in the value column alongside and enter the value of the attribute, in this case ‘emoticon’.

<pc type="emoticon">:-)</pc>

<pc type="emoticon">xoxoxo</pc>

We have not yet used the following tags

Insert tag <INSERT> … </INSERT>

This tag is used to mark up textual or visual material imported into the text from an external source, e.g. a magazine cut out or a photograph.

Note: this is an addition to the LCCW tag set.

Page break tag <PB> … </PB>

This tag is used to mark up any place where a writer deliberately introduces a page break (versus simply writing across two pages). Often this will be signalled with a page number added by the child.

Table tag <TABLE> … </TABLE>, <ROW> … </ROW>, <CELL> … </CELL>

This tag is used to mark up any information presented by the writer in tabular form, that is, information that is intended to be read both horizontally and vertically. It also requires <ROW> and <CELL> tags. These tags will be nested within the overall <TABLE> tags. (See example on p. 221 of Smith, McEnery and Ivanic.)

Appendices

Appendix one

Team members, their roles and their personal identity codes

Name	Identity Code	Roles / responsibilities and tasks
Jeanette King	JK	Project Leader • Maintains the TM Project Manual • Maintains TEI mark-up /tagset • Edits the marked-up texts • Liaises between team members • Organises project tasks and monitors project progress • Organises project funding and budget • Uploads the completed texts to LaBB-CAT • Enters additional demographic information into LaBB-CAT • Analysis of data • Interpretation of data • Project write-up and publications
Christine Brown	CB	Researcher • Liaises with schools and teacher - participants • Organises and files consents and permissions • Collects and codes original scripts, photographs and files the originals • Organises early feedback to the participating teachers • Analysis of data • Interpretation of data
Mary Boyce	MB	Researcher • Drafted the original TM Project Manual. • Devised the original TEI mark-up /tagset • Interpretation of data • Project write-up and publications (e.g. book chapter already solicited)
Mike Davey	MD	transcribes texts (no longer working on the project)
Roberta Tainui	RT	checks transcripts and adds mark-up (no longer working on the project)
Niwa Wehi	NW	transcription and mark-up
Caitlin Swan	CS	transcription and mark-up
Classroom teachers		• Liaise with Christine Brown • Apply results to classroom practice

Appendix two

Decisions regarding the transcription spelling of non-standard items

There are a number of accepted variant spellings of words in Māori (for example, teina and taina). The person doing the markup will not correct these variants but will inform the Project Leader so that they can be added to the LEMMA list below which will record the LEMMA and the word forms that come under each lemma. The list will be periodically sorted into alphabetical order. The LEMMA will be the most frequent spelling from the Corpus of Māori Texts for Children.

Note:
This does not include spellings that are deemed to be errors rather than acceptable variants. Errors will be corrected in the transcript and tagged using the regularisation tag:

LEMMA	Word forms
ēnei	ēnei eneki ngēnei
kei	kei kai
kīa	kīa kīia
māu	māu māhau
mātou	mātou mātau
meneti	meneti miniti
motokā	motokā motukā
netipaoro	netipaoro netipōro
paihikara	pahikara paihikara
tēnei	tēnei teneki
tētahi	tētahi tētehi
wikeni	wikeni wikini

whutupaoro	whutupaoro whutupōro

Appendix three
List of languages and their abbreviations

The two letter ISO639-1 codes, available at
http://www.mathguide.de/info/tools/languagecode.html are used:

Language	**2 letter code**
English	en
French	fra
Māori	mi
Samoan	sm
Spanish	es

Appendix four

TEI Header <TEI HEADER>

Each transcript has a TEI header that stores metadata for that file (e.g. writer information, school code and so on). Transcribers will enter the following information

1. Using keystroke <CTRL>+<ALT>+<H> on your keyboard (or right-click on the Text view and select *Plugins|Form fill* from the popup menu). You should see a form with the header details of the transcript. Fill in as follows. You can use the tab key to quickly move from one field to the next.

Field value	*What this means/refers to*	*example*
title	Title of file, which is participant's code plus the number of piece of writing	00158-20
transcriber name	Your name	Jeanette King
number in series	The number of the piece of writing	20
participant id	The participant's code	00158
age years	The age of the child at time of writing – see instructions below	5
age months		8
school year	The year level that the participant is currently in.	5
creation date	The date the child wrote the piece of writing. If not stated in the text, this may have to be estimated from previous and later pieces of writing	12/7/2013

How to calculate how old the child was at the time they wrote the text.
Once you have worked out the date of the piece of writing (as above) use this date, and the child's date of birth (from excel file) and carefully enter them into the Pearson website:
https://www.pearsonclinical.com.au/agecalculator

We only want the child's year and month (we don't need days) to fill in in the Age sections above. Please fill in in the following way.

If the calculator says
5 years, 2 months, 25 days write 5 in the year field and 2 in the months field (that is, don't round up to 3 months)

8 years, 0 months, 3 days write 8 in the years field and 0 in the months field.

Appendix five

Summary list of tags used in the Tuhinga Māhorahora project

Based on the Appendix (page 225) of: Smith, N. , T. McEnery and R. Ivanic. 1998. 'Issues in Transcribing a Corpus of Children's Handwritten Projects.' *Literary and Linguistic Computing*, Vol. 13 No. 4, 1998. pp. 217-225.

Tag	**Purpose**
<ab>	'Anonymous block'. Separates chunks of text, e.g. dates, paragraphs, headings.
<figure>	Graphic element produced by child.
<name>	Two uses: to faciltiate anonymisation, where necessary and to ensure all names can be excluded from analysis.
<note>	Note of any kind, e.g. abrupt end to a text.
<quote>	To indicate speech or writing which have been quoted from an external source.
<choice><orig><reg>	Regularized spelling, including macrons. Child's original spellings are preserved within the `orig` tag.
<sic>	For unknown words produced by the child.
<unclear>	To note any text which is hard to read or decipher.
<pc>	To tag emoticons

As yet unused tags

<insert>	To indicate textual or visual material from an external source, for example, a photo or something cut out from a magazine.
<pb>	Deliberate page division used by writer.
<TABLE>	Any clear use of table, i.e. tabular data intended to be read both horizontally and vertically.