Developing a Part-Of-Speech tagger for te reo Māori

Aoife Finn
Te Hiku Media
aoife@tehiku.co.nz

Peter-Lucas Jones
Te Hiku Media
peterlucas@tehiku.co.nz

Keoni Mahelona
Te Hiku Media
keoni@tehiku.co.nz

Suzanne Duncan
Te Hiku Media
suzanne@tehiku.co.nz

Gianna Leoni
Te Hiku Media
gianna@tehiku.co.nz

Abstract

This abstract discusses the development of a Part-of-Speech tagger for te reo Māori, which is the Indigenous language of Aotearoa - New Zealand. It mostly focuses on the creation of a tagset that is appropriate for Māori. This is in consideration of the fact that some tagsets have existing tags that are not suitable for some Māori word classes. Alternatively, the existing tagsets might lack entirely a suitable tag for some Māori word classes. And finally, some existing tagsets do not adequately reflect a Māori worldview. Emphasis is put on the importance of capturing the language according to the conceptualization of its speakers, and not imposing "traditional" grammatical categories where it is not appropriate. The solution involved changing how some existing tags are used and in some cases creating entirely new tags that are appropriate for Te reo Māori. The Part-of-Speech tagger was successfully built by a Māori Indigenous organisation and is being used as the foundation for other applications.

1. Introduction

This paper discusses the development of a Part-of-Speech tagger for Māori called *Whakairo Kupu*, meaning to *carve* or *sculpt words*. It specifically focuses on the creation of a tagset that was appropriate for Māori. Our current precision and recall scores are approximately 93%. Hereinafter, Māori will be referred to as te reo Māori or alternatively just Māori, and Universal Dependencies will be abbreviated to UD. Part-of-speech will be abbreviated as POS.

Furthermore in this paper, linguistic examples will consist of four to five lines. The first line will include a morpheme by morpheme te reo Māori phrase or sentence. The second line will see each morpheme with a linguistic gloss that gives information about the syntactic properties or meaning of the morpheme. A third line will show the POS tags that our tagger would assign to the morpheme. The fourth and typically final line will show the English translation. However, in the uncommon instance that a literal translation is needed, it will be shown on the fifth line. For an example of this, please see (1).

1)

Example		Glosses
Haere	mai	*te reo*
go	DIR	*linguistic*
VERB	MOD	*POS*
"Welcome!"		*translation*
Lit: "Go hither"		*literal translation*

Moving on, te reo Māori is the Indigenous language of Aotearoa, also known as New Zealand, (Morrison, 2011). It is a member of the Eastern Polynesian branch of the Austronesian language family which itself has approximately 1200 members, (Harlow, 2007). Māori is related to other members of the Polynesian branch of Austronesian such as Rapanui, Rarotongan, Tahitian, Tuamotuan, Marquesan, Hawai'ian and Mangarevan, (Du Feu, 1996). Te reo Māori is a head-first and dependent- marking language, it is analytical with a high degree of polysemy.

Prior to the development of this tagger, there was no POS tagger for Māori from Aotearoa. POS taggers tag words according to their syntactic or grammatical category. However, many traditional syntactic categories, and by consequence POS labels, do not "work for" Māori, see (2). By this we mean for some of the traditional categories:

2)

a) The definition of, or guidelines for, an existing category is not suitable for Māori.

b) They do not have an existing category for certain word classes of Māori.

c) They do not reflect a Māori worldview of the Māori language.

We wanted a tagset that is usable with industry-wide tools, but we also needed a tagset that would meet the needs of te reo Māori. After researching various tagsets, we decided to base our tagset and guidelines on the UD tagset and tagging conventions. However, the categorization of words has been significantly altered to be appropriate for Māori. This is because at the time of development of our POS tagger, the UD conventions had still not been used to tag a Polynesian language such as te reo Māori, nor did it provide any guidelines about how to tag them.

Therefore the question arose as to how do we tag these words. Of course, we looked at how languages, other than the “big languages” such as English, were tagged. Yet, what works for other languages does not necessarily work for Māori. Furthermore, it would be a mistake to presume that the tagging solution for one Indigenous language should be applied to all Indigenous languages. As part of the re-Indigenization and decolonization, we do not homogenise Indigenous languages.

At this point, it is fitting to take a moment to digress and remind ourselves that at Te Hiku media our vision statement is *He reo tuku iho, he reo ora* which means *A living language transmitted intergenerationally*. This vision statement informs every decision that we make at every level. That means that it is of the utmost importance that we faithfully and accurately capture te reo Māori, as the language that has been passed down intergenerationally. In the same vein, we do not want to impose grammatical categories that are not correct or applicable.

To that end, we worked with highly-proficient, specially-selected Māori speakers and linguists who are specialists in Māori. This has ensured that our POS labels and guidelines conventions faithfully reflect a Māori speaker’s conceptualization of their language.

We achieved this by simply asking speakers. We elicited answers without using questions that were influenced by academic theories of language or pedagogical methods of language teaching. The speakers reviewed our guidelines on a regular and consistent basis, they also partook in a survey to target special areas of interest. Furthermore, our guidelines are evergreen, meaning that they can and do change based on speaker feedback. This does not impact negatively on our tagged corpora as we have an automation system in place to retag words when necessary. We now briefly explore each point above in (2) seriatim.

2. Existing categories are not suitable for te reo Māori.

As mentioned above, some existing definitions and their guidelines for both syntactic categories and POS labels are not suitable for Māori.

The UD conventions follow a lexical approach, that is one-word equals one-tag. However, as mentioned previously, Māori is a highly analytic language in the sense that there are many words with multiple grammatical functions, as opposed to inflection. Sometimes a single concept is represented by many lexical words, see (3). Therefore we worked with our speakers to see when and where single or multiple labels were appropriate.

3)

Māori and POS label(s)				English
Kei te				*present*
AUX				*tense*
Mōku				*for me*
ADPRON				
He	aha	ai		*why*
AUX	PRON	PART		
I	te	rā	nei	*today*
ADP	DET	NOUN	DET	

4) Ignoring white space between written words, in your mind is "i te rā nei"...

a) Made up of a single word “i te rā nei”

b) Made up of many separate words, "i", "te" and "rā" and “nei”

c) Other, please elaborate

We achieved this by asking non-leading questions. For example, in order to establish if the words of *i te rā nei*, meaning *today*, should receive a single or separate tags, we asked questions such as that in (4). If speakers had answered (a), then we could infer that *i te rā nei* should receive a single tag. On the other hand, if our speakers had answered (b), then the words should be tagged separately. We also left a blank space in (c) to allow our speakers to provide any other suggestions. As it happened, for time phrase adverbials with many lexical words such as *i te rā nei*, our speakers overwhelmingly chose to tag each word separately.

Crucially though, this was not the case for all concepts that were represented by many lexical words, as our speakers indicated that certain types should be tagged with a single word. As such, by working with our speakers we avoided making a blanket judgement and were able to give single or separate tags when and where appropriate, all according to the conceptualization of te reo Māori by speakers. Some developers of tagging guidelines for other languages choose a blanket approach for this type of problem. For example in the POS tagging of Griko, all apostrophes between words are treated as a single token, (Anastasopoulos et al, 2018). However this was not the right approach for us or te reo Māori, as evidenced by the fact that our speakers chose both single word and separate word tagging.

5)	Kua	hoko-na	e	au
	PFV	buy-PASS	ADP	1SG
	AUX	VERB	ADP	PRON

he	whare
DET	house
DET	NOUN

"A house has been bought "

6)	Kua	whā	tau	au	ki
	PFV	four	year	1SG	ADP
	AUX	NUM	NOUN	PRON	ADP

Aotearoa
Aotearoa
PROPN

"I have been in Aotearoa for four years"
Lit: "Have been four years, I in Aotearoa"

Moving on, in example (5) tense is marked on the verb *hoko* with *kua*. The token *hoko* is given the POS tag VERB, and the separate tense-marker token *kua* is given the POS tag AUX. However, tense and aspect can also be marked on numbers in Māori, Harlow (2015: 256). This is the case in example (6) wherein *whā*, or *four*, is also marked with the perfect aspect marker *kua*. This is in the same way that verbs, such as *hoko* in (5) are tense-marked. This is not limited to te reo Māōri, numbers that behave like verbs are also found in Choctaw and Jarawara (Dixon, 2012).

Whilst acknowledging that a number can be an "determiner, adjective or pronoun". The UD guidelines do not provide for numerals that behave like verbs. Yet, they state that verbs are often associated with "tense, mood" and "aspect". Therefore, under UD tagset guidelines, these numbers would likely be labelled as VERB.

Notwithstanding, tagging in this way would not be an accurate representation of te reo Māori. So as the POS gloss in (6) shows, we do not adhere to this. The tense-marked number token *whā* is tagged as numeral/NUM. Whilst, the separate tense-marker token *kua* is tagged as AUX .

3. Categories for certain word classes of Māori does not exist.

As stated above, UD conventions sometimes do not have a suitable existing category for certain classes of Māori words. Ergo, we have added POS labels that faithfully capture Māori, both the grammatical categories and the Māori view of te reo.

Māori has a word class commonly known as "particles" in linguistic literature, Harlow (2007: 24). These particles are small words such as *anō, iho, noa, pū, tonu* etc. Each particle can have meaning and many grammatical functions. Following our own analysis of over ninety particles, we found that grammatically they served many purposes, that their syntactic behaviour is wide, varied and commonplace. As such they do not fall under the remit of any "traditional" grammatical categories

For example, the "particle" *rawa* can modify nouns, pronouns, verbs, adjectives, numerals and negatives, (Harlow, 2015). We show a selection of these below. In example (7), *rawa* modifies the pronoun *koutou*. *Rawa* can also modify verbs like *hangaia* in (8), confirmation that verbal modification is taking place can be gleaned from the passive agreement that takes place on *rawatia*. The adjective *wera* is modified by *rawa* in (9). Whereas (10) and (11), show *rawa* modifying a negative and question word, i.e. *kāore* and *aha*, respectively.

7)	Mā	koutou	rawa	e
	ADP	3PL	MOD	TNS
	ADP	PRON	MOD	AUX

rangatira	te	kōrero
lead	DET.SG	discussion
VERB	DET	NOUN

"It is you who should lead the discussion"

8)

Hanga-ia	rawa-tia	he
build-PASS	MOD-PASS	DET.INDF
VERB	MOD	DET

whare	hou	mōna
house	new	ADP.3SG
NOUN	ADJ	ADPRON

"A new house was built especially for her"

9)

He	wera	rawa	te
PRED	hot	MOD	DET.SG
AUX	ADJ	MOD	DET

kai?
food
NOUN

"Is the food too hot?"

10)

Kāore	rawa	mātou	i	mōhio
NEG	MOD	3PL	PST	know
PART	MOD	PRON	AUX	VERB

"We really do not know"

11)

He	aha	rawa	te
PRED	what	MOD	DET.SG
AUX	PRON	MOD	DET

take?
reason
NOUN

"What is the reason?"

Of course, it is fair to ask why we did not use the UD POS tag "Particle", hereafter PART, for te reo Māori "particles". As per the UD guidelines, PART is said to often encode grammatical categories such as "negation, mood, tense", see UD guidelines, (References section below). However, crucially the "particles" of te reo Māori do not encode any of these categories. The UD PART tag is also a landing spot for words "that do not satisfy definitions of other universal parts of speech". For Indigenous or non-European languages, such as Māori, this in particular feels unsatisfactory. Rather than providing an accurate tag, anything that is deemed to fall outside of "universal" grammar is cast-off into the ambiguous PART category. Therefore, we chose to create a POS tag that would be fitting for this part of te reo Māori grammar. In a wider context, this fits with our vision statement mentioned above.

It should be noted however, that when and where the UD PART tag was applicable it was used and does appear in our tagset. This is the case for all the UD tags, we did not create new tags just for the sake of it. An example of the PART tag being used in our data is with te reo Māori words of negation, such as *kāore* in (10).

12)

Kāore	au	i	haere
NEG	1SG	PST	go
PART	PRON	AUX	VERB

"I did not go"

There is another class of words for which there is no suitable traditional label. When first-person singular, second-person singular and third-person singular pronouns, i.e. *ahau, koe* and *ia*, combine with certain adpositions, i.e. t*ā, ā, tō, ō, mā* and *mō* they combine into a single word, (Bauer, 1997). These new combinations are concurrently both pronouns and adpositions. This can be seen in example (14) wherein *tō* and *ahau* have combined into *tōku*. By contrast, *tō* does not combine with *koutou* in (13).

13)

Me	hoki	au	ki
DEON	go_back	1SG	ADP
AUX	VERB	PRON	ADP

tō	koutou	whare
SG.POSS	3PL	house
ADP	PRON	NOUN

"I should go back to your house"

14)

Me	hoki	au	ki
DEON	go_back	1SG	ADP
AUX	VERB	PRON	ADP

tōku	whare
SG.POSS.1SG	house
ADPRON	NOUN

"I should go back go to my house"

These are not very common, but do occur in other languages, such as Irish, where they are commonly called *prepositional pronouns*. A UD Tagset that was developed for Irish simply tags these as preposition/PREP. Yet, this representation is not as accurate as it could be, they are at once both prepositions and pronouns in the grammar of Irish. Furthermore, the UD guidelines do not provide for such a word class.

With this in mind, we worked with our Māori speakers and linguists to faithfully capture and represent the equivalent te reo Māori word class. From working with our Māori speakers and linguists, it became clear that UD conventions do not have a suitable label for either "particles" or "adposition-pronouns". As such we created two new Māori specific labels for our tagset, i.e. modifier/MOD and adposition-pronoun/ADPRON.

4. Categories do not reflect a Māori worldview of the Māori language.

As has been said above, some UD conventions do not reflect a Māori worldview of the Māori language. For instance, the term Māori indicates Indigenous to Aotearoa. By contrast, *Pākehā* means of European origin, and *te reo Pākehā* is the Māori term for the English language. In our corpus, there are some instances of code-switching between Māori and English, and also between Māori and other Polynesian languages.

The UD guidelines recommend that foreign words receive the POS label "X", however this is problematic for us. Although the English language is not Indigenous to Aotearoa, to label English language words as "X" fails to capture the complex bi-cultural reality of modern-day Aotearoa. And to label other Polynesian languages as foreign disregards the historical, linguistic, cultural and genealogical ties among Pacific peoples. If we were to use "X" to tag all words that are not in te reo Māori, then English and other Polynesian languages would be conglomerated, or homogenised, into one group. Furthermore, it also limits the usefulness of our tagger for future applications where these languages are often mixed.

This resulted in the creation of two further Māori specific labels, Pākehā/PAKEHA for English language words, and MOANANUI for the cousin-languages of Māori. The creation of these Pākehā/PAKEHA and MOANANUI labels, allow us to distinguish other languages from te reo Māori, without disregarding the connections between the speakers of te reo Māori and other Polynesian languages.

The UD guidelines and tagsets have been used to tag languages where there is code-switching such as Turkish-German and Frisian-Dutch. It is our understanding that in such cases both languages are given UD tags. This approach would not work for us for two reasons. Firstly, as a small Māori Indigenous organisation, POS tagging English would not be a worthwhile use of our resources. Secondly, while we need to differentiate the other Polynesian languages from te reo Māori in our data, we would not create a tagset, nor presume to tag them without permission from the speakers of those languages.

In summation, the words in our Māori corpora have been categorised and labelled to reflect Māori in the minds of its speakers. At present, this same Māori lead approach is being expanded to include a feature layer that would include features relevant to Māori such as kupu mino and te reo ā-kāinga which are similar but different to loanwords and dialect respectively. Even at the most surface level of our tagging conventions, we do not use terms like dialect, when they are not appropriate to Maōri society.

5. Conclusion

Our tagset uses a total of 21 POS labels. They have been used to annotate our datasets, which contain over 40,000 tokens. The datasets cover many genres and are being constantly expanded. We have used our tagset and annotated datasets to build *Whakairo Kupu*, our POS tagger for te reo Māori. In our most recent *Whakairo Kupu* model, the precision was 92.5%, and the recall was 93.1%. These increased from 86.3% and 48.3% respectively in the very first model.

With regard to sharing our data, or allowing the use of Whakairo Kupu, Te Hiku Media operates under its *Kaitiakitanga* Licence. This quotation in (15) from our P*apa Reo* website best explains it. For more about the *Kaitiakitanga* Licence see our *Papa Reo* website (References section below).

15) Te Hiku Media have developed a Kaitiakitanga licence, which states that data is not owned but as cared for… Te Hiku Media are merely caretakers of the data and seek to ensure that all decisions made about the use of that data respect it's mana and that of the people from whom it descends…Māori data will not be openly released, but requests for access to the data, or for the use of the tools developed under the platform, will be managed using tikanga Māori.

In terms of applications for *Whakairo Kupu*, as it stands, not only does it POS tag te reo, but it has been used to build a grammar checker. It is also being used as a foundation for building a Named Entity Recognition tagger for te reo Māori.

6. References

Anastasopoulos, Antonis and Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, David Chiang. 2018. *Part-of-Speech Tagging on an Endangered Language: a Parallel Griko-Italian Resource*. In Proceedings of the 27th

International Conference on Computational Linguistics, pages 2529–2539 Santa Fe, New Mexico, USA.

Bauer, Winifred. William Parker Te Kareongawai Evans, Te Aroha Noti Teepa. 1997. *The Reed Reference Grammar of Māori*. Auckland: Reed Books.

Dixon, R.M.W. *Basic Linguistic Theory 3*. Oxford: Oxford University Press.

Du Feu, V. 1996. Rapanui. London: Routledge

Harlow, Ray. 2007. *Māori A Linguistic Introduction*. Cambridge: Cambridge University Press.

Harlow, Ray. 2015. *A Māori Reference Grammar*. Wellington: Huia Publishers.

Morrison, Scotty. 2011. *The Raupō Phrasebook of Modern Māori*. Auckland: Penguin Group NZ

Te Hiku Media ~ PapaReo Kaitiakitanga License https://papareo.nz/#kaitiakitanga

7. Abbreviations

1	first-person	num	number
3	third-person	PART	particle
ADP	adposition	PASS	passive
ADPRON	adpositional -pronoun	PFV	perfect
AUX	auxiliary	PL	plural
DEON	deontic modality	POSS	possessum
DET	determiner	PRED	predicative
dir	directional	PRON	pronoun
INDF	indefinite	PST	past
MOD	modifier	SG	singular
NEG	negative	TNS	tense marker
NOUN	noun	VERB	verb