Annotating "Particles" in Multiword Expressions in te reo Māori for a Part-of-Speech Tagger

Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, Gianna Leoni
Te Hiku Media
1 Melba Street, Kaitaia, Aotearoa - New Zealand
aoife@tehiku.co.nz, peterlucas@tehiku.co.nz, keoni@tehiku.co.nz, suzanne@tehiku.co.nz, gianna@tehiku.co.nz

Abstract

This paper discusses the development of a Part-of-Speech tagger for te reo Māori, which is the Indigenous language of Aotearoa, also known as New Zealand. Te reo Māori is a particularly analytical and polysemic language. A word class called "particles" is introduced, they are small multi-functional words with many meanings, for example *ē, ai, noa, rawa, mai, anō* and *koa*. These "particles" are reflective of the analytical and polysemous nature of te reo Māori. They frequently occur both singularly and also in multiword expressions, including time adverbial phrases. The paper illustrates the challenges that they presented to part-of-speech tagging. It also discusses how we overcome these challenges in a way that is appropriate for te reo Māori, given its status an Indigenous language and history of colonisation. This includes a discussion of the importance of accurately reflecting the conceptualization of te reo Māori. And how this involved making no linguistic presumptions, and of eliciting faithful judgements from speakers, in a way that is uninfluenced by linguistic terminology.

Keywords: Māori, te reo Māori, Part-of-Speech Tagging, Indigenous languages, annotation

1. Introduction

This paper discusses a selection of the multiword expressions that occur in the data used to train a Part-of-Speech tagger for Māori.

Hereinafter, multiword expressions will be referred to as MWEs throughout and Part-of-Speech will be called POS. Whilst Māori will be referred to as te reo Māori or alternatively just Māori. Universal Dependencies will be abbreviated to UD. Unless otherwise stated, in this paper "word" means "orthographic word", i.e. in the written form of the language, a word is separated by white space from other words. By MWE we mean more than two orthographic words that commonly occur and are used today together as a phrase.

We wanted to annotate te reo Māori data and to use it to train a model and achieve our goal of building a POS tagger for te reo Māori. Crucially, while doing so our further goal was to use a tagset that authentically captured te reo Māori. The POS tagger itself, was to be eventually expanded to include a features layer, hereafter FEAT layer, which would add more precise information to the POS labels, for example adding the number and gender of a pronoun. The POS tagger was also to be used as a building block for other natural language processing tools, for example Named Entity Recognition and sentiment analysis etc.

Before proceeding further, it is worth noting that the vision statement of Te Hiku Media is *He reo tuku iho, he reo ora* meaning *A living language transmitted intergenerationally*. This foregrounds the importance of capturing te reo Māori as it truly is, as the language that has passed down through Māori *whānau* (family) from generation to generation. Our mission statement is W*hakatōkia, poipoia kia matomato te reo Māori o ngā haukāinga o Te Hiku o Te Ika* which means *Instil, nurture and proliferate the Māori Language unique to haukāinga of Te Hiku o Te Ika.* This stresses our commitment to the revitalization of te reo Māori to capturing, nurturing and facilitating its growth.

2. A Brief Introduction to te reo Māori

Te reo Māori is the Indigenous language of Aotearoa, which is also known as New Zealand, (Morrison, 2011). It is a member of the Austronesian language family which has approximately 1200 members, (Harlow, 2007). Māori belongs to the Eastern Polynesian branch of Austronesian along with Rapanui, Rarotongan, Tahitian, Tuamotuan, Marquesan, Hawai'ian and Mangarevan, (Du Feu, 1996). According to the Statistics New Zealand government website, there are approximately 185,955 people who registered as speaking Māori in the 2018 census, (see References section below). Māori is a VSO, head-first, dependent-marking language.

Like many of its Polynesian counterparts, Māori is an analytical language, which means that it has many many small words or the aforementioned "particles" that indicate the grammatical roles of words. Some examples of particles include *kē, ai, noa, rawa, mai, anō* and *koa*. This paper will particularly focus on MWEs that consist of these so-called "particles".

Furthermore, Māori makes great use of polysemy. This means that a single word can have many meanings and many uses. To somewhat illustrate the extent of polysemy in te reo Māori, we look at the sentence *i whara tāku waewae i a Mere i te hōpua heoi i tino riri au i a ia* (1), in which *i* appears four times. In this single sentence, the *i* shows both past tense and also location, which would both receive the POS label AUX. It also marks the agent of the neuter verb *whara* and a direct object, which would both receive the POS label ADP. See the POS labels in the third line of gloss.

1.	I	whara	tāku	waewae
	PST	injure	my	leg
	AUX	VERB	ADPRON	NOUN

Proceedings of the 18th Workshop on Multiword Expressions (MWE 2022) n@LREC2022, pages 67–74
Marseille, 25 June 2022. © European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

i a Mere i te
AGT ART Mere LOC DET
ADP PART PROPN AUX DET

heoi tino riri au
so PST tino annoy 1SG
CCONJ AUX MOD VERB PRON

i a ia
DO ART 3SG
ADP PART PRON
“Mere hurt my leg in the pool, so I was very annoyed at her”

If expanding the POS labelling to include a more fine-grained FEAT layer, then the difference between these labels needed for *i* are more striking. The *i* in (1) would receive four different FEAT labels, AUX-pst, AUX-loc, ADP-agt and ADP-do. Outside of this, the word *i* is also used in sentences of comparison. This demonstrates the grammatical variation that a single word can show in a single sentence.

It is also worth emphasising further that neither adjacency nor ordering consistently predict grammatical roles nor how labels should be distributed between words. This can be demonstrated with the identical sentences in (2) and (3). In (2) *kei te* is considered a single word and would receive AUX which is a single POS label, see third line of gloss. It would receive a FEAT label of AUX-pres.

2. Kei te mahi ia
PRES work 3SG
AUX VERB PRON
“She is working”

3. Kei te mahi ia
LOC DET NOUN 3SG
AUX DET NOUN PRON
“She is at work”

On the other hand, in (3), *kei te* is two separate words, *kei* would receive AUX and *te* would receive the DET POS label, see third line of gloss. If including a FEAT layer, *kei* would be labelled AUX-loc and *te* would be labelled with DET-sg.

Therefore, thanks to its particularly analytical and polysemous nature, it can be said that the grammatical role of a word is not always clear or easily ascertained in te reo Māori. Moreover, it is often the case that neither adjacency nor ordering are helpful in this same regard. This complexity of correct labelling of words in te reo Māori presents an obvious challenge to POS tagging, both when annotating and training a model to tag correctly.

Furthermore, as attested by our vision and mission statements mentioned above, our organisation is committed to faithfully and accurately capturing and representing te reo Māori. We do not want to colonise the language with terminology where it is neither applicable nor appropriate, and often founded in European theories of language. In that same vein, our concerns lie with faithfulness to the language, rather than metrics. That is to say that we would rather accurately tag te reo Māori with tags that represent Māori conceptualization, and have initially lower metrics that we can improve on, rather than tagging with an unsuitable tagset and inaccurately representing the language. We view such inaccurate tagging as linguistic colonialism. This is especially pertinent because of the effects of colonisation on the Māori language. So while we did want our annotation guidelines to be compatible with industry standards when possible, it was equally, if not more, important that they had to be appropriate for te reo Māori. Therefore we needed to find a “sweet spot” that best fulfilled both of these criteria.

To begin, the UD guidelines are “based on a lexicalist view of syntax”, see References section below. As such, the UD guidelines strongly encourage what we call a one-word-one-POS-label approach. However, that straight-forward lexicalist approach encouraged by the UD guidelines presents problems for te reo Māori. This is problematic because, as shown above in examples (1) through (3), a single word in te reo Māori can have more than one meaning, and crucially more than one use in the grammar of te reo Māori.

It follows that some of the traditional UD grammatical categories for POS tagging were not suitable for use in te reo Māori. At the time of development of the POS tagger, the UD guidelines had never been used to tag an Eastern Polynesian language such as Māori. In that sense, the word classes of Māori are unprecedented from the point of view of UD guidelines. Therefore, we needed to review the existing UD tagging protocols, assess where they were suitable for te reo and if not, then devise tagging new tagging protocols.

From this careful review and considered pre-examination of the UD tagset, our te reo Māori speaking linguists were able to ascertain that parts of the UD tagset would not be suitable for te reo Māori. Having done so, we did not need to use value time or resources tagging te reo Māori with the existing UD tagset. We are a small Māori Indigenous organisation, and given what that would involve, such as training of annotators etc etc, it would not be a worthwhile use of our resources.

ADJ	adjective	PART	particle
ADP	adposition	PRON	pronoun
ADV	adverb	PROPN	proper noun
AUX	auxiliary	PUNCT	punctuation
CCONJ	coordinating conjunction	SCONJ	subordinating conjunction
DET	determiner	SYM	symbol
INTJ	interjection	VERB	verb
NOUN	noun	X	other
NUM	numeral		

Table 1: Universal Dependencies POS labels

On account of this, our annotation guidelines for the POS tagger for Māori were somewhat based on, although non-identical to, the UD guidelines. The 17 labels of the UD guidelines are shown in Table 1. For more information about their requirements see UD guidelines, (link in References section below).

Of interest in this paper, is that during the development of the POS tagger for Māori and these tagging protocols, the issue of MWEs arose and more specifically, the issue of how they should be annotated.

To reiterate, because it cannot be overstated, keeping in mind the unique grammar and history of te reo Māori, it was paramount that we captured the Māori language as accurately as possible and not impose European ideas on the language where they are neither applicable nor appropriate.. We applied this way of thinking throughout our approach to the grammar of te reo Māori. However, in this paper we will limit ourselves to the examination of the word category from te reo Māori called “particles” and specifically when they occur in MWEs.

3. Single Particles in te reo Māori

Before looking at the “particle” MWEs in te reo Māori, we need to familiarise ourselves with their discrete parts, that is the “particles” themselves.

Te reo Māori has a word category called “punga”, they are also known as particles, (Harlow 2007). Again, they are small words like *anō, iho, noa, pū, tonu*. A single particle can perform many different functions in Māori. Our investigation of ninety particles found that some particles can accompany and modify up to five different word categories amongst the categories of verbs, nouns, pronouns, adjectives, numerals and negatives. Because the particles do not fit the traditional definitions, or indeed UD definitions, such as adjectives and adverbs we cannot say that the grammatical role is known, at least not in a way that falls under “traditional” grammatical roles. Furthermore and perhaps most importantly, Māori linguists themselves, such as (Harlow, 2007) and (Biggs, 1969), do not use traditional labels to refer to this word class. Therefore, we have a word class that is lacking an appropropriate POS label.

Given that the meanings of the particles are so varied and nuanced, we will simply gloss them as their orthographic word form in the examples in this paper.

For example, the particle *tonu* can modify verbs, nouns, adjectives and negatives. This effectively places it in the categories of adverb and adjective at the same time, as well being a modifier of numerals and negatives. In example (4) it modifies the passivised verb *waiatatia*, we can be sure of this because *tonu* has the added suffix *-tia* to match that of the passive verb. That would typically place *tonu* in the grammatical category of adverbs.

4. Kei te   waiata-tia   tonu-tia
PRES   sing-PASS   tonu-PASS
tēnei   waiata
DET   song
“The song is still being sung”

It is worth mentioning that while te reo Māori does make use of some suffixes, such as the passive suffix here, it is not a language that makes use of inflection and so these are rare throughout the grammar and their use is very limited. In example (5) *tonu* modifies the locative noun *roto*, which leaves it behaving more like a typical adjective.

5. Kei   roto   tonu   koe
LOC   inside   tonu   2SG
i   te   whare?
ADP   DET   house
“Are you still inside the house?

For the avoidance of doubt, we know that *tonu* is modifying the words that it succeeds because te reo Māori is head initial, and as such modifiers follow the modified. In (6) *tonu* modifies the number *toru*. Example (7) shows us *tonu* modifying the adjective *whero*. Whilst finally in (8), *tonu* modifies the negative *kāore*.

6. E   toru   tonu
PRED   three   tonu
ngā   āpōrō
DET   apple
“There are still three apples”

7. He   whero   tonu
AUX   red   tonu
te   putiputi
DET   flower
“The flower is still red”

8. Kāore   tonu   te   rangatira
NEG   tonu   DET   chief
i   haina
PST   sign
“The chief did not sign”

These examples serve to illustrate the breadth and variety of grammatical uses of particles, and how even when considered singularly they cannot and should not be categorised under traditional grammatical categories.

Be that as it may, central to our interest here is that particles can combine with other particles to create MWEs. What’s more, particles can also combine with other words that are not particles, and these combinations create entirely new MWEs. In summation, as regards annotation for the POS tagger, we encountered three challenges. Namely;

How should:

- single word particles, such as *tonu* above, be annotated.
- particles when combined with other particles, be annotated.
- particles when combined with other non-particles, be annotated.

4. Particles Combined with Other Particles

We have seen an example of a single word particle above with *tonu*. Yet, as stated previously, particles can combine with other particles. Furthermore, the combined meaning is not always a direct combination of the single particle meaning.

To give an example, *noa* is a single particle that has many subtle and distinct meanings. The meanings of *noa* are often connected to ideas that have been variously translated *as being without restraint, casually, by accident, spontaneously, randomly, without restriction, merely, solely* and *only*.

Similarly to *tonu* and other particles, it is multifunctional in its grammatical uses and can modify verbs, nouns, adjectives, question words and numbers, and negatives. In example (9), *noa* is modifying the verb *pakipaki*. In (10) the noun *meneti* is modified by *noa,* whereas in (11) *noa* modifies the adjective *māmā*.

9.	E	pakipaki	noa
	PROG	clap	noa
	ana	au	
	PROG	1SG	

“I am clapping wildly”

10.	E	rima	meneti	noa
	PRED	five	minute	noa
	hei	wehe	māku	
	SCONJ	leave	for_me	

“I have only 5 minutes to leave”

11.	He	māmā	noa
	AUX	simple	noa
	te	whai	whakaaetanga
	DET	have	agreement
	i	a	rātou
	ADP	ART	3PL

“An agreement can be gained relatively easily for them”

In (12) *noa* modifies the question word *aha*. And finally in (13), the number *kotahi* is modified by *noa*, whereas in (14) the negative *kīhai* is modified by *noa*.

12.	He	aha	noa	te	paku?
	AUX	what	noa	DET	little

Lit: “What is merely the smallness?”
“Why all the fuss?”

13.	Kotahi	noa	te
	one	noa	DET
	teina	o	Te Pairi
	Brother	ADP	Te_Pairi

“Te Pairi had only one brother”

14.	Kīhai	noa	kia	tae	te
	NEG	noa	PREF	arrive	DET
	pukapuka	a	Hōne-Heke		
	Letter	ADP	Hōne-Heke		

“Hone-Heke’s letter had not arrived”

Yet, when *noa* combines with other particles in which case the meanings can shift again. When *noa* is combined with another particle *iho*, the combination usually gives the sense of *just, only, that and nothing better, s*ee Harlow (2015: 93). Example (15) gives this sense of *noa iho* meaning *just*. By itself, the particle *iho* has many meanings and uses but is most often a directional particle meaning *downwards* like in (16).

15.	He	whakaaro	noa	iho
	AUX	idea	noa	iho

“It’s just an idea”

16.	Heke	iho
	Get_off	iho

“Come down”

Noa can also combine with the particle *atu*. The primary function of *atu*, although it is one of many, is that of a directional particle indicating direction away from the speaker. This is the case in (17) wherein it specifies the direction of the verb *haere*.

17.	Haere	atu
	go	atu

“Go away”

18.	He	reka	noa	atu	
	AUX	tasty	noa	atu	
	ngā	tītipi	i	ngā	rare
	DET	chip	ADP	DET	candy

“Chips are much more tasty than candies”

When *noa* joins with *atu* to become *noa atu,* it is used to intensify comparative senses, as in (18) where it intensifies the adjective *reka*. *Noa atu* can also indicate that something happened *a long time ago*, thus it becomes a kind of time adverbial MWE, see (19).

19.	Kua	haere	noa	atu
	PERF	go	noa	atu
	au	ki	Itāria	
	1SG	ADP	Italy	

“I went to Italy a long time ago”

This leads onto another particle combination, that is *noa* with the particle *ake*. By itself, *ake* is primarily another directional particle indicating upward motion, see example (20) in which it specifies the direction of the verb *piki*. Yet, when in combination with *noa,* it has a similar meaning to *noa atu*, i.e. *a long time ago,* see (21) where it is a time adverbial MWE.

20.	E	piki	ake		
	PROG	climb	ake		
	ana	au	i	te	maunga
	PROG	1SG	ADP	DET	mountain

“I’m climbing up the mountain”

21.	I wehe	ia	noa	ake
	PSTleave	3SG	noa	ake

“He left a long time ago”

Concluding this section with particle MWEs, and specifically the final two which can serve as time adverbials, we now move to look at particles with non-particles in time adverbials MWEs.

5. Particles in Time Adverbs

Sometimes, adverbs in Māori are single word expressions such as *inanahi* in example (22). However many adverbs, specifically adverbial phrases of time, consist of many orthographic words and as such are time adverbial MWEs. As seen previously, the particles themselves have many varied uses and meanings and a particular combination can mean a variation in the meaning of the time adverbial MWE.

22.	I	haere	ia	inanahi
	AUX	go	3SG	yesterday

“She went yesterday”

By way of illustration, *muri* is usually a locative noun meaning *back, rear* or *behind*. It is shown used in this way in (23). However, it also provides a base for many time adverbial MWEs. In these time adverbial MWEs, it is accompanied by an adposition, and a combination of particles, of which the number can vary.

23. I muri te ngeru
PST behind DET cat
i te rākau
ADP DET tree”
“The cat was behind the tree”

Basic types of time adverbial MWEs can be seen in (24) and (25). In (24) the particle *i* marks past tense and it is followed by *muri* and the particle *iho,* previously seen in examples (15) and (16). The combined overall meaning of this MWE in (24) means *after*. However, it can be seen that the substitution of *iho* with *mai* in (25) extends the overall meaning to include *later* and *afterwards*.

24. I muri iho
PST back iho
“After”

25. I muri mai
PST back mai
“After, later, afterwards”

The particle *mai* also has many many meanings and uses but, like *iho,* is very often used as a directional particle indicating motion towards the speaker as in (26).

26. Whakarongo mai
Listen mai
“Listen to me”

The addition of more particles can again change the meaning. If *tonu* is added to the sentence *i muri tonu iho,* seen in (24), to become *i muri tonu iho*, then the meaning shifts to *straight after,* see (27). But if *tonu* is replaced with *tata* the meaning again changes, but this time it changes to *soon after* as in (28). *Tata* is another particle with various meanings but it often means something akin to *near, almost, slightl*y or *just*. This is how it is used in (29).

27. I muri tonu iho
PST back tonu iho
“Straight after”

28. I muri tata iho
PST back tata iho
“Soon after”

29. Kua tata maoa te kai
PERF tata cook DET food
“The food is almost cooked”

Tata can also be added to *i muri mai* which again alters the meaning to *shortly after* as in (30). And when *tonu, iho* and *tata* are combined, the meaning transforms again into immediately after as in (31). In addition, if the particle *i* is changed to *ā,* the entire tense shifts from past to future as in (32).

30. I muri tata mai
PST back tata mai
“Shortly after”

31. I muri tata tonu iho
PST back tata tonu iho
“Immediately after”

32. Ā muri atu
FUT back atu
“In the future”

It could be said that these examples really bring into focus that in every language, there can be a discrepancy between an idea and the number of orthographic words.

This can be seen using both te reo Māori and English time adverbial MWEs as examples. In (33), both languages express the idea of “the day after today” with the single words, *āpōpō* and *tomorrow*. In (34), the idea of “the day after the day after today” is expressed in te reo Māori with a single word *ātahirā*, whereas in English it has many words. By contrast, as shown in (35), the idea of “today” is represented with a single word in English, whereas it is a four-word time adverbial MWE *i te rā nei* in te reo Māori.

33. Āpōpō
“tomorrow”

34. Ātahirā
“the day after tomorrow”

35. I te rā nei
ADP DET day DET
“Today”

The dilemma that faced us was how should these time adverbial MWEs be tagged? Should each orthographic word receive a POS label, and if so with which labels? Or should the phrase be tagged as a single unit, and if so with which POS label?

6. Solution

The previous sections looked at a selection of the various ways in which particles can occur in MWEs in the grammar of te reo Māori.

The question arose as to how we decided to tag them, and how we reached those tagging decisions. Repeating earlier sentiments, we strove to both reflect and to capture the conceptualization of te reo Māori that has been handed down from generation to generation. And importantly, to not presume or impose grammatical characteristics where they are neither applicable nor accurate. And as the UD guidelines had not been developed for, nor used with, a POS tagger for an Eastern Polynesian language such as te reo Māori, we needed to devise a way to capture how speakers conceptualised their language. This was mainly achieved by two methods.

We simply set out to work with te reo Māori speakers, in order to establish their conceptualization of their language. Our group of speakers consisted of highly proficient, specially selected te reo Māori speakers and also te reo Māori-speaking linguists. They have been termed our *rangatira reo*, which roughly translates as “esteemed Māori language leaders”. *Rangatira reo* is both the singular and plural term..

We were cognizant of the fact that many speakers' terminology for grammar might have been influenced by their past education, i.e. any pedagogical methods used during their language learning, or any academic theories of the language. These often come with their own terminology. The terminologies, whilst they might be useful for their purpose, are not always the best suited to te reo Māori. A well-known example of this is the verbal category in te reo Māori that are very often known as “stative verbs” in linguistic literature and in learners theory and exercise books. An example of such a verb is shown in (36). Unsurprisingly, in casual conversation many speakers refer to these verbs as “stative verbs”, although upon examination they have proven not to be stative in nature.

36. I pau te kai
PST consume DET food
i te ngeru
ADP DET cat
“The cat ate up the food”

Knowing that this could have had an influence on any feedback we received, we strove to mitigate any influences from the past experiences of our *rangatira reo.* Whatsmore, we ourselves did not want to suggest or mention these terms and to unduly influence their answers to our questions. Bearing this in mind, we set out to elicit responses about te reo, but we did this using non-leading questions free of terminology.

37. In these two sentences: "*kei te haere au ki Te Awamutu i tēnei rā*" and "*kei te haere au ki Te Awamutu āpōpō*", are "*i tēnei rā*" and "*āpōpō*" doing the same thing?

    a) Yes, "*i tēnei rā*" and "*āpōpō*" are doing the same thing in the two sentences

    b) No, "*i tēnei rā*" and "*āpōpō*" are not doing the same thing in the two sentences

    c) Other, please elaborate

    d) Please feel free to add any comments, thoughts or insights about the question above or your answer to it.

We wanted to begin by making no presumptions in our examination of te reo Māori with the *rangatira reo*. To that end, we began by checking the most basic premisses. We believed, but strived to confirm, that the time adverbial MWEs such as *i tēnei rā* and the single word adverbs such as *āpōpo* are in fact doing the same job in a sentence. However, we did not want to use the word *adverb*, lest any preconceived ideas of what an adverb should be influence the answer. So we used a simple question such as that in (37) wherein no grammatical terminology was mentioned.

If the *rangatira reo* answered a) then we could then presume that the purpose of both phrases is adverbial. It turns out that this was the most popular answer and is indeed the case.

Having established the basics, we also took this approach with the particles that combined with other particles. For these simple particle combinations, if the *rangatira reo* had answered with a) in (38), we could reasonably infer that *noa iho* should receive a single POS label. If answer b) had been predominant then it could be deduced *noa* and *iho* should receive separate tags. Should c) have been chosen then we could ascertain that yet again *noa* and *iho* should receive separate tags but in the UD syntactic relations layer, the words would be linked together as a flat MWE. Finally, answer d) serves to provide the *rangatira reo* with the opportunity to share their own thoughts or feedback.

38. Ignoring white space between written words, in a phrase such as "*he whakaaro noa iho*", in your mind, is "*noa iho*"...

    a) Made up of one word "*noa iho*"

    b) Made up of two separate words, in this case "*noa*" and "*iho*"

    c) Made up of two separate words, but they are acting together as one, in this case "*noa*" and "*iho*"

    d) Other, please elaborate

To offer a further example, if phrased in a particular way some questions might prompt a particular response, such as the question in (39). First of all, the question names certain grammatical categories i.e. *verb*, *noun*, *adjectiv*e and *adverb* and therefore could implicitly suggest them as the answer to our question. Secondly, naming certain grammatical categories it presumes that those “traditional” grammatical categories are appropriate for te reo Māori. This is unsatisfactory because it allows the possibility that the true conceptualization of te reo Māori is overlooked, and a POS label that is neither accurate nor appropriate is applied.

39. Is the “*ake*” in “*ā muri ake nei*”...

    a) a verb

    b) a noun

    c) an adjective

    d) an adverb

    e) other
    If other, please specify_____

Therefore, we opted to use questions phrased like those in (40). These non-leading questions do not suggest nor do they presume the appropriateness of grammatical categories. Again, we were very clear in what we were asking and how each answer was to be interpreted.

40. Ignoring white space between written words, is a phrase such as "*Ā muri ake nei*"...

    a) A single word, made up of one phrase "*ā muri ake nei*"

    b) Made up of many separate words, in this case "*ā*", "*muri*", "*ake*" and "*nei*"

    c) Made up of a primary word "*muri*" which is described by other words like "*ake*" and "*nei*"

    d) Other, please elaborate

To illustrate, in (40), if the *rangatira reo* had answered with a), that would mean that the four words *ā muri ake nei* would receive one single POS label. If the *rangatira reo* had answered with b), each word would receive one POS label. However, in the case that the *rangatira reo* has answered with c) then each word would still receive one POS label, but in the UD syntactic relations layer, each word would also be shown as a dependent of the noun *muri*. This option was influenced by and included due to linguistic knowledge that *muri* is typically a noun, and that it is likely that the other words modify it. Its inclusion could be said to be based on a “hunch” from linguists who speak te reo Māori. However, it is important to mention that the inclusion of c) is just that, and if *rangatira reo* had given negative feedback about it, then it would have been immediately discounted. Finally, there is d) to allow for any unanticipated feedback that the *rangatira reo* may have to contribute. Indeed, if they felt it was appropriate, a *rangatira reo* could have used this opportunity to advocate for the use of traditional grammatical categories, or alternative labelling.

As it happens, c) was markedly the most popular answer, followed by b). This affirmed our "hunch" that either way, each word should receive a separate POS label. No *rangatira reo* identified the time adverbial MWE as a), thereupon ruling out a single POS label for the MWEs.

Whilst the questions in (38) and (40) established that the particles in the MWEs should be tagged with separate POS labels. It still needed to be made clear exactly what labels the particles should receive, and if those labels ought to be from the traditional grammatical categories suggested by the UD guidelines. To that end, we used many questions like that in (41). Again, shying away from explicitly using terms like adjective and adverb etc, we tried to use very neutral language, and we were clear about what we were asking and how the answers should be interpreted.

41. In these two sentences: "*i mahi pai koe?*" and "*he kaitaraiwa pai koe*", is "pai" doing the same thing?
    a) Yes, "*pai*" does the same thing in the two sentences
    b) No, "*pai*" does not do the same thing in the two sentences
    c) Other, please elaborate

In (41), two sentences were given, the first sentence *i mahi pai koe* is translated as *you worked well*, with *pai* describing the verb *mahi*, thus behaving like an adverb, see (42). The second sentence is *you are a good driver* with *pai* describing the noun *kaitaraiwa* therefore behaving like an adjective, see (43).

If a *rangatira reo* answered a), it signified that *pai* is performing the same grammatical role in both sentences and so is capable of behaving like both an adjective and an adverb. Therefore it falls outside of any traditional grammatical roles and requires a new POS label. If a a *rangatira reo* answered b), it signified that *pai* is not not behaving in the same way in the sentences and they should receive different POS labels, such as the traditional UD adjective and adverb labels. However, for these types of questions the *rangatira reo* answered a) indicating that there is a single grammatical category in te reo Māori that does not behave like either an adjective or adverb. Rather, it is more fluid and can modify nouns, verbs and many other grammatical categories as seen in the earlier sections of this paper.

42. I mahi pai koe
PST work pai 2SG
"You worked well"

43. He kaitaraiwa pai koe
AUX driver pai 2SG
"You are a good driver"

Therefore, bearing in mind our most important goal to accurately and faithfully capture te reo Māori as it is conceptualised in the minds of speakers, we created a new label for these types of words. This label was for particles, and for now it is called modifier or MOD. To illustrate the labelling protocols drawn from our rangatira reo, a time adverbial MWE like *ā muri ake nei* would be annotated for our datasets with the POS labels shown in the third line of the gloss (44), and eventually our POS tagger would tag it in the same way.

44. Ā muri ake nei
FUT back ake nei
ADP NOUN MOD MOD
"In a little while"

All in all, we asked 151 specially designed questions covering 10 different areas of interest. The UD guidelines have 17 labels. We use all of them, but we have an additional 4 labels, including modifier/MOD, that are for te reo Māori. It is important to spotlight that we did not create labels for the sake of it, rather we created them when they were expressly needed.

Coupled with asking our *rangatira reo* our specially designed questions. We have an ever-changing set of guidelines for our annotators. These guidelines were and are under constant review from the *rangatira reo*, meaning that the guidelines are evergreen. By evergreen we mean that, when required, they can always be changed. The guidelines are not static, so when we receive feedback from our *rangatira reo* that our annotation protocols are no longer appropriate, we alter the guidelines immediately. In essence, this means that while the decolonial and re-indigenizing processes are ongoing, the guidelines are being adapted to reflect the latest and most appropriate POS labels for te reo Māori. Of course, that begs the question that if our guidelines are updated, how can the annotated data also remain up-to-date. The answer is a simple one, we have an automatic tagging system in place, that means any words can be retagged as needed, and the POS tagger can be retrained.

7. Conclusions and Future Work

This paper has discussed the challenges encountered when tagging the MWEs of te reo Māori. Ultimately, our annotated datasets included a total of 21 POS labels. Where appropriate MWEs were annotated with te reo Māori appropriate labels, and they were annotated with the correct number of labels suitable for the conceptualization of that particular MWE.

At the time of writing, our datasets have over 40,000 tokens, with text taken from informal text, formal text and from social media. Most importantly, our datasets have successfully trained a model. As such we have successfully built a POS tagger for te reo Māori, called *Whakairo Kupu* meaning *carver of words*. Our current precision and recall are both at 93%.

In terms of access to both the data and the POS tagger, Te Hiku Media operates under its Kaitiakitanga Licence, see an abridged version in (45). More information about the Kaitiakitanga Licence can be found on our Papa Reo website, see references.

1. Data is not owned but as cared for under the principle of kaitiakitanga and any benefit derived from data flows to the source of the data… 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. Te Hiku Media have been invited to speak on their kaitiakitanga licence and it has been adopted by a government department and a social enterprise.

The POS tagger *Whakairo Kupu* has already been used as a base on which to build a grammar checker for te reo Māori. A FEAT layer is almost complete and it is currently being used to produce a Named Entity Recognition tagger.

8. Abbreviations

1	first person	NOUN	noun
2	second person	PART	particle
3	third person	PASS	passive
ADP	adposition	PERF	perfect
ADPRON	adpositional-pronoun	PL	plural
AGT	agent	PRED	predicative
ART	personal article	PRES	present
AUX	auxiliary	PROG	progressive
CCONJ	coordinating conjunction	PRON	pronoun
DET	determiner	PROPN	proper noun
DO	direct object	PST	past
LOC	location	SCONJ	subordinating conjunction
MOD	modifier	SG	singular

Table 2: Abbreviations

9. References

Biggs, Bruce. 1969. Let's Learn Māori: A Guide to the Study of the Māori Language. Auckland: Reed.

Du Feu, Veronica. 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

Universal Dependencies - Tokenization and Word Segmentation https://universaldependencies.org/u/overview/tokenization.html

2018 Census totals by topic – national highlights https://www.stats.govt.nz/information-releases/2018-census-totals-by-topic-national-highlights-updated