Towards Universal Dependencies in Cook Islands Māori

Sarah Karnes[1] **Rolando Coto-Solano**[1] **Sally Akevai Nicholas**[2]
[1]Department of Linguistics, Dartmouth College
[2]Waipapa Taumata Rau University of Auckland, New Zealand
sarah.r.karnes.23@dartmouth.edu
rolando.a.coto.solano@dartmouth.edu
ake.nicholas@auckland.ac.nz

Abstract

This paper presents a first attempt at applying Universal Dependencies (De Marneffe et al., 2021) to Cook Islands Māori, a Polynesian language from the Cook Islands. There is limited previous work on dependency parsing of Austronesian languages, and there are no existing treebanks for any Polynesian languages. This paper presents a treebank for Cook Islands Māori, with the goal of creating a syntactic resource for further research as well as language instruction. We also present a list of structures that have proved challenging or difficult to parse (e.g. negative sentences, equative sentences, spatial prepositional phrases, actor emphatic structures, and stative sentences with an agent). The treebank contains 126 sentences with 1035 tokens taken from Nicholas (2017), and annotated using UD v2 guidelines. We parsed the sentences using a context-free grammar and then generated dependency parses automatically by using the head-floating method of Xia and Palmer (2001). We are working to expand this treebank and use it to annotate existing corpora and create pedagogical tools for the community. Finally, we review some of the data sovereignty issues related to using Indigenous language data in large language models

1 Introduction

This paper presents the process of creation of a Universal Dependencies (De Marneffe et al., 2021) treebank for the Cook Islands Māori language (henceforth CIM). CIM is a Polynesian language, closely related but distinct from languages like te reo Māori, Samoan, Tongan and Tahitian. UD treebanks exist for other Austronesian languages such as Cebuano, Indonesian and Tagalog, but no previous work has been done on dependency parsing for the Polynesian branch of the Austronesian family.

CIM is spoken by approx 2,500 people in the Cook Islands (Ministry of Finance and Economic Management, Government of the Cook Islands, 2021), plus an additional 10,000 in Aotearoa New Zealand and Australia (Nicholas, 2018). Amongst the diaspora in New Zealand only 9% speak the language (Statistics New Zealand, 2018), and within the Cook Islands themselves familial transmission has decreased, especially on the more populated islands of Rarotonga and Aitutaki (Nicholas, 2018).

CIM is a predominantly isolating language, but it does present derivational morphology (e.g. *'aka-* 'causative' + *rongo* 'hear' = *'akarongo* "to cause to hear"; *tae* 'to reach a place' + *-'anga* 'nominalization' = *tae'anga* "reaching (of the place)"). The unmarked word order in CIM sentences is VSO, or more broadly, predicate initial, as many CIM sentences are nominal. CIM has several distinct dialects; this paper focuses on Southern CIM, spoken on the islands of Rarotonga, Aitutaki, Atiu, Ma'uke, Miti'aro and Mangaia.

There are no existing treebanks for Polynesian languages. There are a few for Austronesian languages, including Tagalog (Aquino and de Leon, 2020), Cebuano (Aranes, 2022) and Indonesian (Alfina et al., 2019). There are a number of Universal Dependency treebanks available for Indigenous languages. They exist for languages of the Americas like Yupik (Chen et al., 2020; Park et al., 2021), Maya Kiché (Tyers and Henderson, 2021), Bribri (Coto-Solano et al., 2021), Shipibo-Konibo (Vasquez et al., 2018), Guaraní (Thomas, 2019) and several other Tupí languages from Brazil (Ferraz Gerardi et al., 2021). They also exist for Australian languages like Warlpiri (Nivre et al., 2020), and Indigenous languages from Eurasia such as Sami (de Lhoneux et al., 2017) and Yakut (Merzhevich and Gerardi, 2022). In this paper we seek to start the creation of UD treebanks for Polynesian languages.

2 Methodology and Data Collection

We parsed 126 CIM sentences, with a total of 1035 tokens. On average the sentences were 8.2 tokens long, with a minimum of two tokens and a maximum of 21 tokens. The data consists of sentences from Nicholas (2017), which includes both isolated sentences from previously published material such as Buse's Dictionary of CIM (Buse and Taringa, 1995), as well as transcriptions of oral narrations gathered through linguistic fieldwork. Most of the sentences come from the Rarotonga and Ma'uke dialects. We used a two step process to parse these sentences: (1) First, we constructed a context-free grammar (CFG) to parse sentences that were manually tokenized and pre-labelled for part of speech. The CFG contains 163 rules: 7 for CPs, 40 for sentences, 15 for VPs, 32 for NPs, 44 for terminals, and 25 for other non-terminal structures. (2) With the CFG parse as a base, we used the method of Xia and Palmer (2001) to raise the heads of the CFG subtrees and establish dependencies between words. This part of our code also converts the XPOS used in the Nicholas (2017) linguistic description into UPOS and sets UD v2 relations between words.

Figure 1 shows an example CFG parse and its corresponding dependency structure. It shows elements that are common to other Polynesian languages. For example, the verb is preceded and followed by tense-aspect-mood (TAM) markers, which are labeled as auxiliaries.

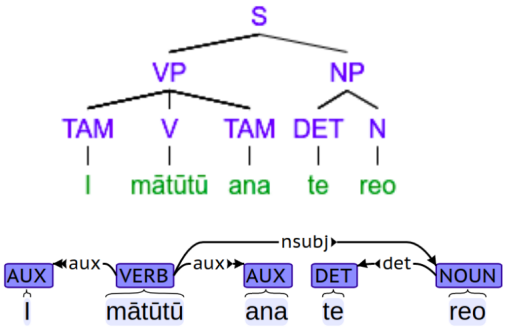

Figure 1: CFG and UD parses for *I mātūtū ana te reo* "The language was strong"

3 Results

Table 1 shows the UPOS in the existing dataset. The most common parts of speech are NOUN, DET, AUX and VERB, accounting for 61% of the tokens in the treebank.

NOUN	192 (19%)	ADV	72 (7%)
DET	149 (14%)	PROPN	46 (4%)
AUX	141 (14%)	PART	38 (4%)
VERB	139 (14%)	ADJ	18 (2%)
ADP	126 (12%)	PUNCT	14 (1%)
PRON	87 (8%)	Others	13 (1%)

Table 1: UPOS tags in CIM sentences.

Table 2 shows the relations used in the UD treebank. The most frequent relations are case, aux, nsubj, det and root, and they account for 65% of the labels in the dataset.

case	146 (14%)	nmod:poss	34 (3%)
aux	141 (14%)	nmod	28 (3%)
nsubj	130 (13%)	amod	26 (3%)
det	128 (12%)	det:poss	21 (2%)
root	126 (12%)	obl:agt	20 (2%)
advmod	62 (6%)	nummod	18 (2%)
obl	49 (5%)	punct	14 (1%)
dobj	46 (4%)	Others	46 (4%)

Table 2: Relations in CIM parsings

4 Challenging Structures

This treebank is a work in progress, but several issues have been observed which will also be found in other Polynesian languages.

4.1 Negative Sentences

The negative structures in CIM are different from those in most UD languages. Figure 2 has the word *kāre* "not" as the first in the structure. It might be tempting to tag this as a negative adverb, and have the verb *moe* "sleep" as the root. However, multiple analyses of Polynesian languages (Hohepa, 1969; Nicholas, 2017) have shown that *kāre* is actually a verb (diachronically the combination of the TAM *ka* and the adverb *kore* "not"). This should be the root of the structure, and have the second verb as its clausal complement (xcomp).

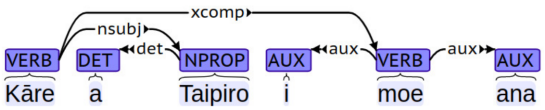

Figure 2: UD parse for the negative sentence "Taipiro didn't sleep"

4.2 Equative Sentences

Figure 3 shows examples of equative sentences. CIM does not have a copula, and therefore these are parsed similarly to their equivalent structure in languages like Russian: The predicate of the sentence, be it a noun or an adjective, is selected as the root. In figure 3a, the name *Mere* is the root, and it is marked as the predicate by the specifying word *ko*. This word "marks a phrase as nominal and specific" (Nicholas, 2017, 188). It usually marks the predicate of the equative, but it can also mark focused arguments

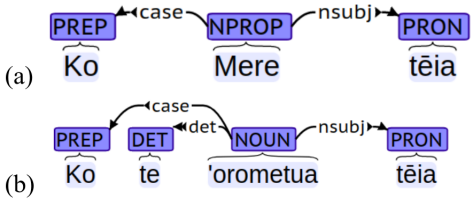

Figure 3: UD parse for (a) *Ko Mere tēia* "This is Mere" and (b) *Ko te 'orometua tēia* "This is the minister"

Given that this is a nominal construction, where the main element of *ko te 'orometua* in figure 3b is the noun *'orometua* "minister", and that there are no TAM markers in the sentence, there are no words that should be labeled as a VERB. The word *ko* is a function word, but it should not be labeled as a PART or an AUX (copula) because it doesn't have any TAM functions. There is research that has analyzed this as a preposition (Brown and Koch, 2016; Massam et al., 2006). If *ko* is a preposition, the relationship between *Mere* and *ko* would be case. This is a very desirable property; this is parallel to languages like Polish, where the predicate of an equative structure needs a particular case (e.g. the instrumental in the case of Polish).

4.3 Spatial prepositional phrases

Figure 4 shows the phrase *ki runga i tōna pona* "on her dress". The word *runga* means the surface of an object, and the preposition *ki* indicates that something is done in/on somewhere. The third word, *i*, usually works as a preposition for place (e.g. *i te kainga* "at home"). But, what is the role of the word *i* in the following parse?

This could be parsed in two ways. The first one is to see this as a sequence of prepositional phrases, whose literal translation would be "on the surface on her dress", and whose root would be *runga*. The second parse shows a potential, multi-

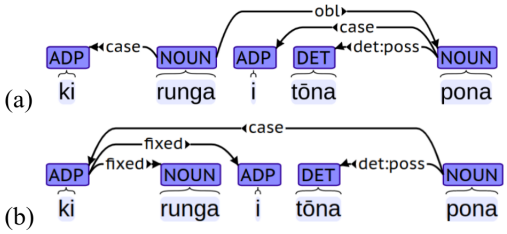

Figure 4: Two possible parses for *ki runga i tōna pona* "on her dress". The first one will be preferred.

word preposition *ki runga i* "on", whose root is the noun *pona* "dress". The literal translation of the second structure would be "on-top her dress". In our ongoing work we have chosen to use the first parse in order to keep the prepositions separate and gain in generalization.

4.4 Actor emphatic structures

CIM and other East Polynesian languages have a particular construction used for specifying the agent of a presupposed transitive event. Polynesianists have designated this the actor (or agent) emphatic construction (Nicholas, 2017, 239). Example 1 shows a typical example.

(1) *Nā Mere i tunu i tēia māngа.*
AEC mere TAM cook ACC DET food

"Mere cooked this food" or "It was Mere who cooked this food".

There are competing analyses for how this construction should be parsed, including differences in how many clauses it comprises, and the grammatical relationships between the constituent phrases (cf. Nicholas, 2017, ch7). Nicholas (2017, 259) analyses the agent phrase as the predicate (root) and the verb phrase as a subordinate clause. The status of the patient phrase is potentially unclear but, in examples such as 1 above, it is the object of the verb. We take it that these constructions are analogous to simple possessive constructions. In figure 5, the person *Mere* possesses the food, and therefore the sentence is translated as "The food is Mere's". In figure 6, the thing that *Mere* possesses is the action of "that [she] cooked the food".

The possessive *nā* phrase and the verb phrase function as two different clauses. The possessive phrase contains the root, just as it does in figure 5. The word *nā* is labeled here as an ADP, as its function is to express the grammatical/semantic relationship of the Nā phrase to the verb.

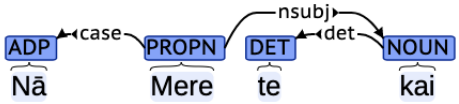

Figure 5: A possible parse for *nā Mere te kai* "The food is Mere's".

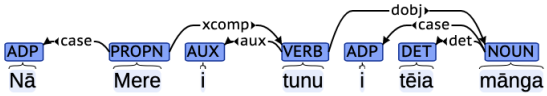

Figure 6: A possible parse for *nā Mere i tuna i tēia mānga* "Mere cooked this food for me".

4.5 Stative sentences with an agent

CIM has a subcategory of verbs commonly called *stative* that typically takes a theme as the subject and optionally includes an agent or cause NP introduced by the preposition *i*. The examples in 2 illustrate this.

(2) a. *Kua pou te taro*
TAM be-consumed DET taro

"The taro is all gone".

b. *Kua pou te taro i te puaka*
TAM be-consumed DET taro AGT DET pig

"The pigs ate all the taro" or "The taro has been consumed by the pigs".

c. *Kua pou te taro i te puaka te kai*
TAM be-consumed DET taro AGT DET pig DET eat

"The pigs ate all the taro by eating" or "By eating, the taro has been consumed by the pigs".

The sentence 2b is stative, but it has the same syntactic structure as a VSO active sentence. Furthermore, it has the same orthographic and phonetic characteristics as a regular transitive sentence. We predict that it will be very difficult for the parser to process these types of sentences without a deeper understanding of their semantics. This structure is only possible with a special lexical category of stative verbs. There are about 100 of these stative verbs, which do not take the -Cia passive suffix and which tend to have theme subjects. The system would need to learn to distinguish these verbs during the training process.

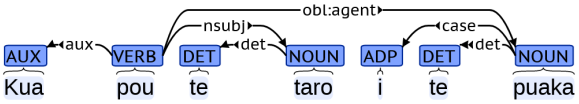

Figure 7: Parse for *Kua pou te taro i te puaka* "The pigs ate all the taro".

A further difficulty presents itself in examples like 2c. In the sentence "The pigs consumed all the taro by eating", the final clause *te kai* "by eating" can be analyzed in two different ways. The word *kai* could be analyzed as a bare VP within a verbal dependent clause, or it could be a bare nominal phrase akin to a gerund (Nicholas, 2017, 309).

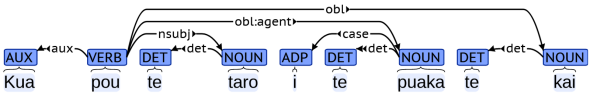

Figure 8: Parse for *Kua pou te taro i te puaka te kai* "The pigs consumed all the taro by eating".

We analyze this sentence in the manner demonstrated in figure 8. There, the word *kai* "food, eating" is marked as a nominal gerund, and is connected to the verb by the obl relation. While this captures the meaning of the structure, it will be a challenge for the parser because the majority of NPs without a preposition are subjects, and would be tagged as nsubj.

5 Data Sovereignty and Future Applications

One major consideration during this project is data sovereignty. As we make the treebank we are considering making it available through the Kaitiaki License from Te Hiku Media (2023). In this license, the mana (property, authority) of the data remains with the population of the Cook Islands, and the data itself is under the care of a Cook Islands researcher (author Nicholas). This means that, if the data were to be used by non Cook Islanders, for example to train a large-scale language model, the people seeking to use the treebank would need to ask permission from the Cook Islands author, as well as seek ways to contribute to the community. The Cook Islands author has a long history of work within her community, and has previously used corpora to create pedagogical tools. The treebank described here is intended to automate and accelerate the annotation of the corpus, but its use in other

tools would be subject to approval by the Cook Islands researcher.

One of the main challenges during the process is to understand how contributors might potentially object to a usage of their data. We have solely used previously published materials for this treebank, but many of the sentences in Nicholas (2017) were uttered by Cook Islanders in direct interviews. We need to consider two components of this work: (i) the treebank itself, and (ii) the potential files where the trained parsing models (e.g. UDPipe models) are stored. In the treebank the people who uttered the sentences are credited, and it is easier for speakers of CIM to either approve the inclusion of their sentence, or to object to their data being part of the database. However, in the case of the model files, the data has been cut off from those who have uttered it. By the time the model receives these parsed sentences as input, the sentences are separated from their identifying metadata, and they become part of the numerical parameters of the model. Explaining this process to people is complex, and it makes it difficult for speakers to potentially object to certain ways of using their data (e.g. to augment a large-scale parsing database). We hope to use the aforementioned license to authorize projects which can have a positive impact on the community, and thereby allay potential objections of community members. However these relationships must be tended continuously by ongoing consultation, collaboration and responsiveness. We will communicate our results to the Cook Islands community through outreach at the University of the South Pacific, and through the connections of the research with the community.

There are two main goals of this work. The most immediate one is to create a fully parsed corpus for CIM, so that the grammatical structure of the language can be further studied. This corpus has been collected by the Cook Islands author, and has been used for creative pedagogical materials. The parser will accelerate the work of studying the grammar of CIM. We also have a further pedagogical goal: We are working to produce a web-based interface for the parser, so that school teachers and students of CIM can get grammatical information to assist them in their studies. Teachers currently take classes at the University of the South Pacific, where they study basic CIM grammar and parts-of-speech is one of their most challenging topics. We will train a POS tagger from the treebank and use this to help teachers in their training, and for them to become more involved in CIM NLP. The parsing model can also be used to build tools like question answering systems and chatbots, which would create a positive impact (Galla, 2016) and ultimately garner attention to the language, particularly from younger community members who might be drawn into the language through technology.

These difficult questions are of course not just part of the work with CIM, but should be incorporated in all NLP work with Indigenous languages. The challenge of balancing community benefit with data openness remains crucial to the progress of NLP for under-resourced languages.

6 Conclusions

In this paper we have presented the first steps towards the creation of a Universal Dependencies Treebank in the Cook Islands Māori language. This would be the first attempt to create such a treebank for a Polynesian language. We have tagged a total of 1035 words and we have plans to continue to expand it before releasing it to the public. We plan to release it under a license that privileges Cook Islander's decisions of how the data will be used. The overarching goal is to use this treebank to accelerate the tagging of CIM corpora and to create pedagogical tools to train teachers in CIM grammar.

7 Acknowledgments

Meitaki ma'ata ki te au tangata tei tuku mai tā rātou reo ki a mātou. Mē kāre kōtou kāre tēia 'anga'anga. Thank you so much to all the CIM speakers who shared their words with us and continue to collaborate in this project. Finally, this project is supported by the Marsden Fund Council from the Aotearoa New Zealand Government (21-MAU-018), managed by The Royal Society Te Apārangi.

References

Ika Alfina, Arawinda Dinakaramani, Mohamad Ivan Fanany, and Heru Suhartanto. 2019. A Gold Standard Dependency Treebank for Indonesian. In *the Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33)*.

Angelina Aquino and Franz de Leon. 2020. Parsing in the absence of related languages: Evaluating low-resource dependency parsers on Tagalog. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 8–15.

Glyd Aranes. 2022. The GJA Cebuano Treebank: Creating a Cebuano Universal Dependencies Treebank. Master's thesis, Itä-Suomen yliopisto.

Jason Brown and Karsten Koch. 2016. Focus and change in Polynesian languages. *Australian Journal of Linguistics*, 36(3):304–349.

Jasper Buse and Raututi Taringa. 1995. *Cook Islands Maori Dictionary*, volume 123. Editorial of University of the South Pacific.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved Finite-State Morphological Analysis for St. Lawrence Island Yupik using Paradigm Function Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2676–2684.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards Universal Dependencies for Bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2021. TuDet: Tupían Dependency Treebank (version v0.2).

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

Pat Hohepa. 1969. Not in English and Eehara in Maori. *Te Reo*, 12:1–33.

Miryam de Lhoneux, Yan Shao, Ali Basirat, Eliyahu Kiperwasser, Sara Stymne, Yoav Goldberg, and Joakim Nivre. 2017. From Raw Text to Universal Dependencies - Look, No Tags! In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 207–217.

Diane Massam, Josephine Lee, and Nicholas Rolle. 2006. Still a Preposition: The Category of Ko. *Te Reo*, 49.

Tatiana Merzhevich and Fabrício Ferraz Gerardi. 2022. Introducing Yakutoolkit. Yakut Treebank and Morphological Analyzer. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 185–188.

Ministry of Finance and Economic Management, Government of the Cook Islands. 2021. Census 2021: Key findings. https://www.mfem.gov.ck/statistics/census-and-surveys/census/267-census-2021.

Sally Akevai Te Namu Nicholas. 2017. Ko te Karāma o te Reo Māori o te Pae Tonga o Te Kuki Airani: A Grammar of Southern Cook Islands Māori. *Doctoral Dissertation, University of Auckland*.

Sally Akevai Te Namu Nicholas. 2018. Language contexts: Te Reo Māori o te Pae Tonga o te Kuki Airani also known as Southern Cook Islands Māori. *Language Documentation and Description*, 15:64.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Hyunji Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142.

Statistics New Zealand. 2018. 2018 Census totals by topic – national highlights (updated). (StatsNZWebsite).

Te Hiku Media. 2023. Kaitiakitanga License. https://github.com/TeHikuMedia/Kaitiakitanga-License.

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)*, pages 70–77.

Francis Tyers and Robert Henderson. 2021. A Corpus of K'iche' Annotated for Morphosyntactic Structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-Konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161.

Fei Xia and Martha Palmer. 2001. Converting dependency structures to phrase structures. Technical report, Pennsylvania Univ Philadelphia.