Crosslingual Dependency Parsing of Hawaiian and Cook Islands Māori using Universal Dependencies

Gabriel H. Gilbert[1], **Rolando Coto-Solano**[2], **Sally Akevai Nicholas**[3], **Lauren Houchens**[2], **Sabrina Barton**[2], **Trinity Pryor**[2]

[1] University of Chicago, [2] Dartmouth College
[3] The University of Auckland (Waipapa Taumata Rau)

Correspondence: ghgilbert@uchicago.edu, rolando.a.coto.solano@dartmouth.edu, ake.nicholas@auckland.ac.nz

Abstract

This paper presents the first Universal Dependency (UD) treebank for ʻŌlelo Hawaiʻi (Hawaiian). We discuss some of the difficulties in describing Hawaiian grammar using UD, and train models for automatic parsing. We also combined this data with UD parses from another Eastern Polynesian language, Cook Islands Māori, to train a crosslingual Polynesian parser using UDPipe2. The crosslingual parser produced a statistically significant improvement of 2.4% in the labeled attachment score (LAS) when parsing Hawaiian, and this improvement didn't produce a negative impact in the LAS of Cook Islands Māori. We will use this parser to accelerate the linguistic documentation of Hawaiian.

Hōʻuluʻulu Manaʻo

I kēia pepa, hōʻike mākou i waihona pepeke mua loa no ka ʻŌlelo Hawaiʻi i ke ʻano Universal Dependency (UD). Wehewehe aku mākou i nā pilikia me ka hoʻohana ʻana iā UD, a waiho mai i kumu mīkini hou no ke kuhikuhi ʻano huaʻōlelo ʻana i ke ʻano hana nona iho. Hoʻohui i ia ʻike me nā palapala ʻōlelo o ke ʻano UD mai iā ʻŌlelo Kuke ʻAilani, a hoʻomaʻamaʻa i mīkini kuhikuhi ʻano huaʻōlelo me UDPipe2. He kōkua maoli nō, me ka maikaʻi aʻe o 2.4% i ka "labeled attachment score" (LAS) me ka ʻŌlelo Hawaiʻi, ʻaʻole naʻe i hoʻopilikia i ka LAS o ʻŌlelo Kuke ʻAilani. Makemake mākou e hoʻohana aku i ia no ka pono o ka hoʻōla ʻōlelo ʻia ʻana o ka ʻŌlelo Hawaiʻi.

1 Introduction

This paper presents the first attempt to construct a Universal Dependencies (Nivre et al., 2020) treebank for ʻŌlelo Hawaiʻi (hereafter: Hawaiian). Hawaiian is an Indigenous Polynesian language spoken in Hawaiʻi as a community language by Kānaka Maoli (Native Hawaiians) and non-Hawaiian residents (Kimura, 1983). Hawaiian has been the subject of intense revitalization efforts since the Hawaiian Cultural Renaissance of the 1970s (Kamanā and Wilson, 2001). Given the need for increased grammatical analysis to further revitalization goals, NLP tools like parsing can potentially facilitate the creation of annotated corpora.

Here we describe the structure of the treebank and use the treebank of a related Polynesian language (Cook Islands Māori) to build a crosslingual Polynesian UD parser.

1.1 NLP for Polynesian Languages

There are two motivations for this project. The first one is to create a parser for Hawaiian in order to conduct syntactic analysis of sentences gathered in the process of linguistic documentation and description. The second, larger goal, is to foster networks of collaboration across linguists from Polynesia, and to join efforts in accelerating documentation of their languages, with the ultimate purpose of language revitalization and normalization. In the case of this project, the Hawaiian author (Gilbert) is collaborating with the Cook Islands author (Nicholas) because Cook Islands Māori is the only other Polynesian language that has a treebank available (Karnes et al., 2023). These two languages are Eastern Polynesian and they are closely related. Their syntax shows numerous commonalities: VSO order, a verbal complex with tense-aspect-mood markers and directionals, a very similar system of articles, demonstratives and numerals, and numerous cognates with relatively transparent changes between the proto-language and the two languages (e.g. *taŋata "person" > CIM *tangata*, Haw. *kanaka*) (Elbert and Pukui, 1979; Nicholas, 2017). These two languages also share the status of being under-resourced in terms of NLP data. We hope to leverage their linguistic commonalities to improve the parsing models and to work towards the common goal of describing their syntax to facilitate language teaching and transmission.

Proceedings of the 18th International Conference on Parsing Technologies (IWPT, SyntaxFest 2025), pages 40–50
August 26, 2025 ©2025 Association for Computational Linguistics

There is previous work on NLP for Eastern Polynesian languages, including work on automatic speech recognition (ASR) for te Reo Māori from Aotearoa New Zealand (Jones et al., 2023) and Cook Islands Māori (Coto-Solano et al., 2022a), development of text-to-speech for both languages (Keith, 2024; James et al., 2024), part-of-speech tagging for both languages (Finn et al., 2022; Coto-Solano et al., 2018), and forced alignment for Cook Islands Māori (Nicholas and Coto-Solano, 2019; Coto-Solano et al., 2022b).

NLP work on Hawaiian has involved forays into speech recognition (Chaparala et al., 2024), morphological analysis (Hosoda, 2019) and orthographic reconstruction (Shillingford and Parker Jones, 2018). Additionally, some large language models like ChatGPT (OpenAI, 2022) and Gemini (GeminiTeam et al., 2024) support the generation and translation of Hawaiian, while speech recognition models like Whisper (Radford et al., 2023) can provide support for Hawaiian ASR. We hope to expand the availability of Hawaiian NLP applications and develop tools to further empower language documentation work.

This project is similar to work on other treebanks for under-resourced languages (Rodríguez et al., 2022; Tyers and Henderson, 2021; Coto-Solano et al., 2021; Ramsurrun et al., 2024). Our main goal is to create treebanks that can help with linguistic documentation. A secondary goal is to use the data collected along the way to create temporary models that can be used for bootstrapping, so that the annotation process can switch from completely manual annotation to computer-aided annotation. By doing this, the model can provide a first-pass of the parsing and a human expert can correct this. This accelerates the process and leads to faster annotation.

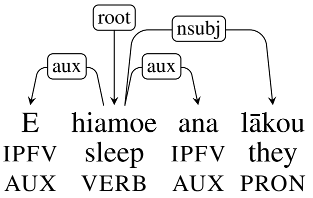
2 Methodology

We will first describe the structure of the treebanks, and then describe the experiments to parse the Hawaiian and CIM data using zero-shot, monolingual and crosslingual methods. In these experiments we attempt to leverage the similarities between Hawaiian and CIM to improve the performance of the Hawaiian model: the model with the least amount of data. We also attempt to leverage a high-resource language, English, to investigate whether its models can aid in the initial parsing of these two low-resource languages.

2.1 Data Sources and Annotation

The first step was to create a dependency treebank for Hawaiian. We collected a total of 145 sentences, containing a total of 1015 tokens. Sentences were 7±3 tokens on average, and came from past documentary linguistic work (12 from Elbert and Pukui 1979; 2 from Pukui and Elbert 1986; 10 from Gilbert 2023), published interviews (24 from Kanahele 1970), and from fieldwork with Hawaiian speakers (98). Sentences were manually annotated using Universal Dependencies 2 (UD2) (Nivre et al., 2020). Example (a) shows a typical parse; the Tense-Aspect-Modality (TAM) particles surrounding the verb root are tagged as auxiliaries, and the subject follows the verbal complex.

(a)



“They were / are / will be sleeping"

We manually tagged the corpus using Universal Parts of Speech (Nivre et al., 2020). Table 1 shows the distribution of POS tags for Hawaiian. The most frequent part of speech is VERB (15%), followed by PUNCT (15%) and NOUN (12%).

VERB	157 (15%)	PRON	96 (9%)
PUNCT	147 (15%)	DET	85 (8%)
NOUN	119 (12%)	PROPN	30 (3%)
ADP	115 (11%)	ADJ	12 (1%)
AUX	112 (11%)	CCONJ	10 (1%)
ADV	107 (11%)	Others	25 (2%)

Table 1: Frequency of UPOS tags in the Hawaiian Treebank (145 sentences, 1015 tokens).

We also annotated the corpus for relations. Table 2 shows that the most common relations in the Hawaiian corpus are root (14%), punct (14%) and case (11%).

Next, we expanded the pre-existing treebank for Cook Islands Māori (CIM) (Karnes et al., 2023). We grew the previous treebank, which contained 126 sentences (1035 tokens), by adding more sentences from a grammar of the language (Nicholas, 2017) and an L2 learning manual (Turepu Carpenter and Beaumont, 1995), manually annotating them using UD2. This new corpus has 663 sentences, with a total of 7658 tokens and an average

root	145 (14%)	advmod	46 (5%)
punct	145 (14%)	compound	39 (4%)
case	115 (11%)	obj	33 (3%)
nsubj	106 (10%)	obl	21 (2%)
aux	96 (9%)	cc	20 (2%)
det	74 (7%)	Others	175 (17%)

Table 2: Frequency of relations in the Hawaiian treebank (145 sentences, 1015 tokens).

sentence length is 12±7 tokens. Table 3 shows the most common parts of speech. The three most frequent ones are NOUN (18%), ADP for adpositions (15%) and DET determiners (14%). All of these occur at higher proportions than in the Hawaiian corpus.

NOUN	1418 (18%)	ADV	513 (7%)
ADP	1122 (15%)	PUNCT	472 (6%)
DET	1094 (14%)	PROPN	239 (3%)
VERB	894 (12%)	PART	155 (2%)
AUX	861 (11%)	ADJ	140 (2%)
PRON	587 (8%)	Others	163 (2%)

Table 3: Frequency of UPOS tags in the CIM treebank (663 sentences, 7658 tokens).

The CIM dataset was also tagged for relations; the summary of these is shown on Table 4. The proportion of nsubj, aux and obj tags in CIM is similar to those in Hawaiian, but the CIM data has more instances of case (15%) and det (12%).

case	1156 (15%)	advmod	446 (6%)
det	953 (12%)	obl	409 (5%)
aux	850 (11%)	nmod	354 (5%)
nsubj	694 (9%)	obj	340 (4%)
root	663 (9%)	amod	161 (2%)
punct	471 (6%)	Others	1161 (15%)

Table 4: Frequency of relations in the CIM treebank (`nmod` includes the possessive `nmod:poss`) (663 sentences, 7658 tokens).

2.2 Zero-Shot and Monolingual Experiments

The first step in our experiment was to train monolingual parsing models for each language. The total number of sentences for each language were randomly split into 80%-10%-10% train/dev/test sets. The test sets belong to the same domain as the training sets: linguistic examples and language learning textbook examples (see section 2.1). We repeated this process 30 times, obtaining 30 unique test sets for each language. The Hawaiian sets had 115/15/15 sentences, while the CIM sets had 531/66/66 sentences. We trained 30 separate models with these sets for each language using the UDPipe2 parser (Straka, 2018). For each model, we calculated the F1 of the Universal Parts of Speech (UPOS), unlabeled attachment score (UAS), and labeled attachment score (LAS). We report the median score of these 30 measures.

Next, we used the monolingual models to test zero-shot parsing with a closely related language. We parsed the original 30 test sets for Hawaiian with the CIM monolingual model. We also parsed the original 30 test sets of CIM using the Hawaiian monolingual model. We evaluated these parses with respect to median UPOS, UAS, and LAS.

One of our ultimate goals in this project is to study the parsing of extremely low-resource Indigenous languages, for which entirely new datasets might need to be built from scratch at great expense to community members, language practitioners, and researchers. If existing models can facilitate this work, we could obtain a considerable head start in new projects. To test this, our next experiment was to parse the Hawaiian and CIM sentences using a zero-shot method, with a model from a completely unrelated language. We chose the `en_core_web_sm` English model from spaCY (Honnibal et al., 2020) because of its easy usability by other researchers. We used this model to parse the same 30 test sets for Hawaiian and 30 test sets for CIM, and report UPOS, UAS, and LAS.

2.3 Crosslingual Experiments

In the second stage of our experiments we trained models where we combined the Hawaiian and CIM data during the training stage. We trained UDPipe2 models, combining the 30 training/dev sets for both languages, and parsed 30 test sets for each language. We also report UPOS, UAS, and LAS for these models.

We then performed an additional experiment where we modified one language to resemble the other. Hawaiian and CIM are closely related, and their cognates show well-attested regular sound correspondences that go all the way back to Proto-Polynesian. Table 5 shows five sound correspondences that are stable enough that they can very transform a CIM word into a Hawaiian word. For example, the CIM word *rātou* 'they' may be changed into its Hawaiian cognate `lākou` by chang-

CIM	Hawaiian
k	‘ (‘okina)
t	k
v	w
' (saltillo)	h
ng / n	n

Table 5: Examples of regular sound alternations between CIM and Hawaiian (Otsuka, 2005).

ing the ‘r’ for an ‘l’ and the ‘t‘ for a ‘k’. These transformations, based on well-documented diachronic processes (Otsuka, 2005), were applied to the CIM data so that it would bear an even closer resemblance to the Hawaiian data. We performed these changes to the 30 train/dev sets of CIM, combined them with the original train/dev Hawaiian sets, and then tested on the Hawaiian test sets.

Finally, we replicated the modified condition, this time modifying the Hawaiian text to more closely resemble the CIM text. For example, the orthography of the Hawaiian word *kākou* ‘everyone’ was transformed into *tātou*, again using the historical sound correspondences in table 5. We applied the first four changes but were unable to do so for the fifth change (n>ng), because the ⟨n⟩ in Hawaiian can be related to either an ⟨n⟩ in CIM (e.g. Haw: *manu*, CIM: *manu* ‘bird’) or to an /ŋ/ (e.g. Haw: *mauna*, CIM: *maunga* ‘mountain’). We applied the one-to-one changes to the Hawaiian sentences and combined them with the original CIM sets. We ran the 30 training rounds and evaluated on the 30 CIM test sets.

In summary, the experiments with the parsing models have five conditions: (i) zero-shot evaluation on an English model, (ii) zero-shot evaluation on a closely-related Polynesian language, (iii) monolingual training, (iv) crosslingual training with data from both Hawaiian and CIM, and (v) crosslingual training where one of the Polynesian languages was modified to more closely resemble the other.

3 Results

Table 6 shows the medians for each language, condition, and metric. Figure 1 summarizes the results for zero-shot parsing versus parsing with a monolingual model trained specifically for each language. Figure 2 summarizes the results for the crosslingual training compared to using monolingual models.

3.1 Hawaiian Models

First, we study the relationship between the zero-shot parses by using an ANOVA model with the zero-shot and the monolingual conditions as independent variables. When zero-shot parsing is performed with a model from a closely related language, it provides significantly better results than when the model is trained on a genetically unrelated language. The zero-shot parses for Hawaiian using an English model have a median of LAS=3%; this is much lower than the parses using a CIM-only model, which have LAS=42% (F(2,87)=1308, p<0.0001; Bonferroni-corrected p<0.0001). These improvements also hold for the other metrics: zero-shot UAS is significantly higher for the CIM-only model than for the English model (66% versus 24%, F(2,87)=691, p<0.0001; Bonferroni-corrected p<0.0001), and zero-shot UPOS follows this pattern (56% versus 19%, F(2,87)=1947, p<0.0001; Bonferroni-corrected p<0.0001).

Using the same ANOVA models, we will study the relationship between the zero-shot parses and the parses generated with the monolingual Hawaiian model. For the three metrics (UPOS, UAS, LAS), the model trained on monolingual Hawaiian data has a significantly higher F1 than the best performing zero-shot approach. When parsing Hawaiian sentences, the monolingual Hawaiian model has a median LAS of 69.5% compared to 42% for zero-shot using a CIM model. UAS has a median of 80.5% compared to 66% for zero-shot with CIM, and UPOS has a median of 85.8% for the monolingual Hawaiian model, compared to 56% when Hawaiian is parsed with the zero-shot using CIM. All of these differences are significant (Bonferroni-corrected p<0.0001).

Finally, we will study the effects of building a crosslingual model by training on both Hawaiian and CIM data. A repeated measures t-test showed that training on both the Hawaiian and CIM data produced a median significant improvement of 1.94% in LAS (t(29)=2.4, p<0.05), compared to parsing with a model trained only on Hawaiian data. When the data is not paired, the difference is larger: LAS_{Cross}: 71.9%, LAS_{Mono}: 69.5%; ΔLAS=2.4%. When the model was trained on a combination of the Hawaiian and modified CIM data (see section 2.3), this produced a smaller but still significant improvement of 1.64% (t(29)=2.3, p<0.05) in LAS compared to the parses generated by the monolingual Hawaiian model. The non-

	Hawaiian			CIM		
	UPOS	UAS	LAS	UPOS	UAS	LAS
(1) Zero-Shot (English model)	19.1	24.3	2.8	16.2	20.9	1.8
(2) Zero-shot (with Polynesian model)	56.1	66.2	41.6	47.4	48.2	28.3
(3) Monolingual	85.8	80.5	69.5	**90.9**	**87.0**	**78.0**
(4) Crosslingual (Hawaiian + CIM)	85.5	81.5	**71.9**	90.7	86.9	77.8
(5) Crosslingual (with modified Polynesian lang)	**86.9**	**81.9**	71.7	90.7	86.9	77.5

Table 6: Median F1 for UD parsing. (In condition 2, the Hawaiian data was parsed using a model trained on CIM, and the CIM data was parsed using a model trained on Hawaiian. In condition 5, we modified the Cook Islands data to match Hawaiian orthography and tested on Hawaiian, and viceversa for CIM).

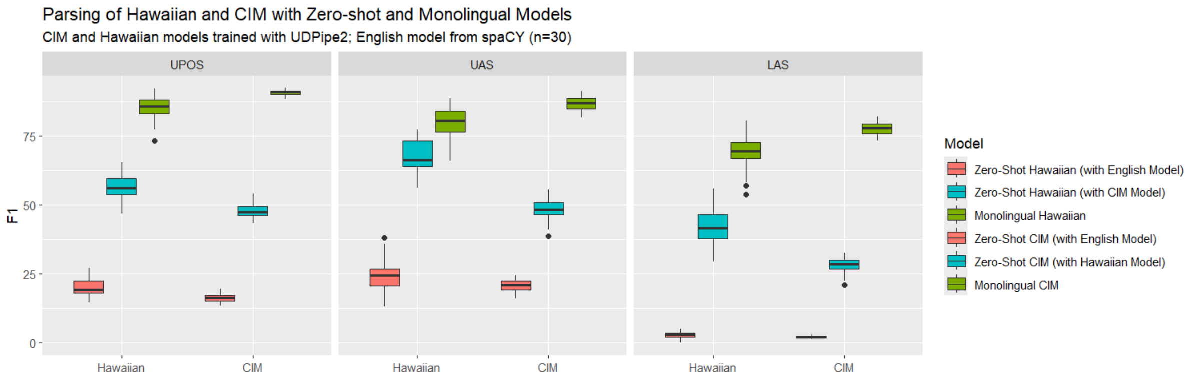


Figure 1: Zero-shot and monolingual parsing for Hawaiian and Cook Islands Māori.

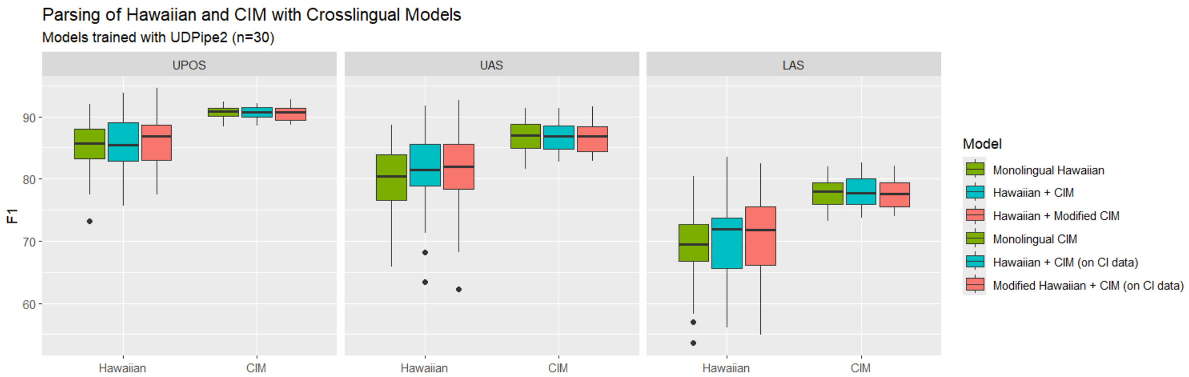


Figure 2: Monolingual and Crosslingual parsing for Hawaiian and Cook Islands Māori.

paired difference is ΔLAS=2.2%.

Both of these patterns also hold for UAS. The crosslingual model had a median paired improvement of 1.7% (t(29)=2.1, p<0.05), and the crosslingual model with the modified CIM data had a median paired improvement of 1.6% (t(29)=1.8, p<0.05). However, this pattern does not hold for UPOS. Training on both the Hawaiian and the CIM data does not provide statistical improvements for UPOS, regardless of whether the CIM data is modified to more closely resemble Hawaiian (ΔLAS_{Modif}=1.1%, p=0.10) or not (ΔLAS_{Modif}=-0.3%, p=0.28).

In summary, the crosslingual training provided a small but significant boost to the parsing of Hawaiian data. Modifying the CIM data did not provide further improvement. Zero-shot parsing is possible, but is improved by using a model from a closely-related language.

3.2 CIM Models

We also studied the performance of zero-shot parsing of CIM (using both English and Hawaiian trained models), monolingual parsing, and crosslingual parsing with the added Hawaiian data. The F1 is lowest when parsing with an English model (median LAS: 2%), compared to parsing with a model trained on the Hawaiian data from section 2.1 (median LAS: 28%, F(2,87)=8411, p<0.0001, Bonferroni corrected: p<0.0001). This holds true for the other metrics: UAS is 21% for the zero-shot English and 48% for zero-shot using the Hawaiian monolingual model (Bonferroni corrected: p<0.0001). This is also the case for UPOS: F1 is 16% for zero-shot English, and 47% for parsing with Hawaiian (Bonferroni corrected: p<0.0001).

When we compare the monolingual versus the crosslingual models, the patterns for CIM are different from those in Hawaiian. There is no significant difference in the parsing results when using the crosslingual model, compared to the monolingual model (LAS_{Cross}=77.8, LAS_{Mono}=78.0, paired t-test p=0.09). Likewise, there are no significant differences between the crosslingual model with modified Hawaiian and the monolingual model (LAS_{Modif}=77.5, LAS_{Mono}=78.0, paired t-test p=0.42). This is also true for the other metrics: there are no significant differences when calculating the UAS ($\text{p}_{Cross/Mono}$=0.14, $\text{p}_{Modif/Mono}$=0.58) or the UPOS ($\text{p}_{Cross/Mono}$=0.60, $\text{p}_{Modif/Mono}$=0.81).

In summary, using a crosslingual model to parse the CIM data does not significantly improve or affect the results, compared to using a monolingual CIM model. Zero-shot parsing of CIM is also better when using a model from a closely-related language (i.e. Hawaiian), but the improvement is much less (ΔLAS=26.5%) than what was found when parsing Hawaiian using the CIM model (ΔLAS=39%), probably because there was much less Hawaiian data to contribute to the learning of CIM.

4 Discussions

In this section, we discuss the performance of the monolingual and crosslingual models, the kinds of errors they make when parsing, and consider some issues encountered while constructing the Hawaiian treebank.

4.1 Crosslingual Parsing

Figure 3 shows the change in LAS between the monolingual and crosslingual (without orthographic modification) conditions. There is a significant performance gain when using the crosslingual model on Hawaiian: the median gain was 2.4% (LAS_{Mono}=69.5% versus LAS_{Cross}=71.9%). But these gains were not uniform. As is depicted in Figures 1 and 2, there is considerable variation for the crosslingual conditions. In fact, for some of the 30 test sets, we actually observed a loss in F1. Gains can be as high as 9.6% (LAS_{Mono}=69.1% versus LAS_{Cross}=78.7%), while losses may be as high as 8.6% (LAS_{Mono}=69.8% versus LAS_{Cross}=61.2%). This pattern should be kept in mind when working with such small datasets, especially where the exact sentences used in each of the train/dev/test sets might have major effects on performance down the line.

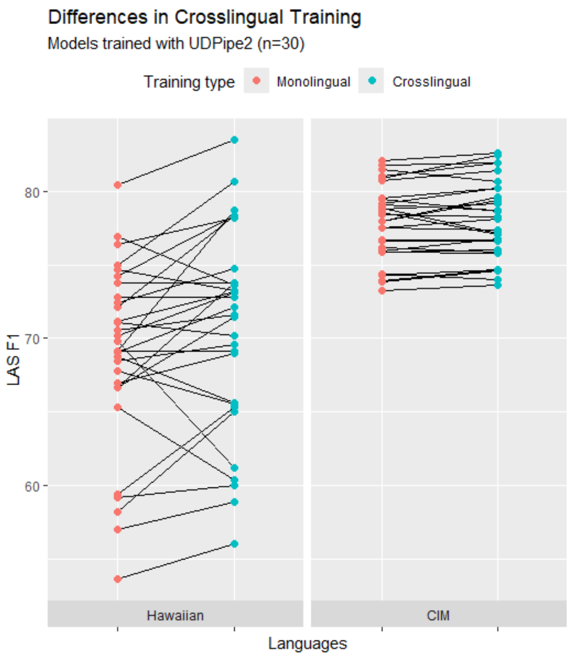


Figure 3: LAS of parses in the crosslingual and monolingual conditions, for specific test sets.

As observed in the results section, crosslingual training did not significantly impact the performance of the CIM LAS. The unpaired difference between the crosslingual and monolingual conditions is 0.2 in favor of the monolingual models, but the paired difference between them is 0.26 in favor of the crosslingual model. Figure 3 shows the values for the 30 test sets, and again we see variation: gains as high as 1.7% in some sets, losses as low as 1.7% in others.

Modifying the orthography of one language to be closer to the other did not provide gains in performance important enough to justify its usage. In the case of Hawaiian, even if the modified dataset did have the highest scores for UPOS and UAS, these were not significantly higher than those of the simple crosslingual model (p=0.36).

In general, the Hawaiian model's performance is higher than that of other similarly-sized models, e.g. Yoruba (140 sentences, UPOS 59, UAS 45, LAS 29, (Dione, 2021)). This is potentially due to Hawaiian's lower number of inflectional suffixes. As for CIM, its performance is comparable to that of models of similar size, for example Ottoman Turkish (9000 sentences, UPOS 88, UAS 62, LAS 52), Tamil (12000 sentences, UPOS 89, UAS 78, LAS 69) and Telugu (6000 sentences, UPOS 94, UAS 91, LAS 84) (Straka, 2025).

4.2 Common Parsing Problems

The total number of errors across all 30 Hawaiian test sets show fewer errors for the crosslingual model than the monolingual model, with respect to both UPOS tagging (421 vs. 414 errors) and LAS (686 vs. 672 errors). The most common problems when tagging relations were mislabeling a coordinating conjunction (cc) as an auxiliary (27 times for the monolingual model vs. 26 times for the crosslingual), labeling a adverbial modifier (advmod) as the root (21 vs. 17), and mislabeling an oblique argument as the direct object (18 vs. 14). As for the parts-of-speech, the most common errors relating to parts-of-speech involved mislabeling adverbs as verbs (34 vs. 39), adverbs as nouns (22 vs. 19), and adverbs as auxiliaries (20 vs. 10).

As for the CIM data, the most common relations errors were mislabeling oblique arguments as objects (141 errors in the 30 test sets parsed using monolingual models), objects as obliques (101 errors), and auxiliaries as case markers (95 errors). These errors possibly stem from the fact that a given word with the form *i* can be either a TAM marker, direct object marker, or locative/temporal oblique marker; the system may still be learning to correctly identify each homophone. As for parts-of-speech, the most common errors were mislabeling pronouns as determiners (123), verbs as nouns (122), and auxiliaries as adpositions (86).

4.3 Challenging Structures in Hawaiian

Here we will discuss three challenges during the construction of the Hawaiian treebank: (i) issues with orthography, (ii) morphology and tokenization, and (iii) dependent clauses.

4.3.1 Orthography

Orthographic representations of Hawaiian offer challenges to straightforward data processing. Hawaiian has a relatively small phonemic inventory: eight consonants (/p/, /m/, /w/, /n/, /l/, /k/, / /ʔ/, /h/) and five vowels (/i/, /e/, /a/, /o/, /u/) with contrastive vowel length (e.g. /pipi/ 'cow' vs. /pīpī/ 'stingy'). Hawaiian's original orthography did not mark glottal stops or vowel length; a standardized orthography developed in the 1970s introduced (1) the *‘okina* for the glottal stop /ʔ/, e.g. */paʔina/* as ⟨pa‘ina⟩, and (2) the *kahakō* (macron) for vowel length, e.g. /laːkou/ as ⟨lākou⟩ (Wilson, 1981).

Most Hawaiian texts predate this modern orthography, and involve both homographs (e.g. *pa‘ina* 'crack' and *paina* 'pine' are both written ⟨paina⟩) and true homophones (e.g. directional adverb *mai* 'towards speaker' vs. preposition *mai* 'since, from'). In this dataset, we included a small number of sentences (three) which were written in both the traditional and the modern orthography, in an attempt to familiarize the system with this variation.

Furthermore, historical and modern orthographic representations often differ in how they write high frequency collocations (e.g. either as a single or as multiple words). Some of these are given below:

Traditional	Modern	Gloss
akula	*aku lā / akula*	'away'
apau	*ā pau / āpau*	'all'
oia	*‘o ia / ‘oia*	'3SG'

Table 7: Orthographic comparisons of collocations.

While decomposition seems possible, these collocations suggest a degree of lexicalization beyond orthographic choice. With the 3SG pronoun, for example, both morphemes may be found elsewhere—*‘o* is a focus marker, *ia* is a demonstrative element meaning 'that (one)'—but inflection of the 3SG requires both. Several parses (three) were created for different variations of the same sentence to enrich the treebank with orthographic variation.

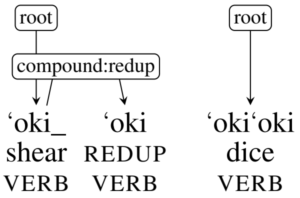
For Hawaiian varieties of slightly different phonological inventories whose speakers may not follow standardized spelling conventions (e.g. the Ni‘ihau community, see NeSmith (2019)), we chose to keep their words as represented by the community, in order to familiarize the treebank with intra-linguistic variation.

4.3.2 Morphology and Tokenization

Hawaiian's particular morphology raised several questions as to the best route for tokenization. Whereas morphological inflection's marginality and non-productivity motivated a simple analysis of morphological processes, the presence of non-concatenative morphology (e.g. vowel lengthening, reduplication) presented a more complex situation that merits comment.

Inflectional morphology surfaces as vowel lengthening (e.g. *wahine* ‘woman’ vs. *wāhine* ‘women’) and reduplication, both partial and total (e.g. *mele* ‘sing’ vs. *memele* ‘sing (pl.)’; *‘oki* ‘shear, cut once’ vs. *‘oki‘oki* ‘dice, cut repeatedly’). Vowel lengthening is reserved for a closed class of roots, and only represented orthographically; items were represented with the plural feature only when rendered with modern orthography. Because reduplication is no longer a productive process, and because certain idiosyncratic meanings were specific to certain roots, reduplicated forms were kept as single morphemes for parsing purposes. Example (b) shows two possible parses for this; the second option was ultimately chosen.

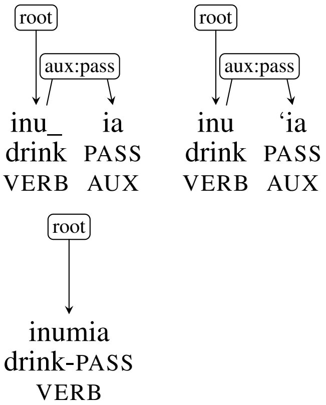
(b)



Two possible parses for:
“to dice, to cut repeatedly"

All prefixes are universally represented as root-attached in texts; we treated prefixed verbs as single items accordingly. There was more variation for suffixes, exemplified particularly by the case of Hawaiian's passive morphemes. There are two types of passives: suffixes {-‘ia} and {-Cia}. In modern texts, {-‘ia} is usually written as a separate word (e.g. *inu ‘ia* ‘[for something] to be drunk’), but in older texts it may appear as a bound suffix on the verb (e.g. *inuia*). We chose to represent -*‘ia* as a separate unit linked by aux:pass to the verb, choosing to align older documents with contemporary norms as in (c). Conversely, {-Cia} is a fossilized passive no longer productive in Hawaiian, limited to just a few specific roots (e.g. *inumia* ‘[for something] to be drunk’). Since the morpheme itself is fossilized and its phonological shape depends on a given root, we chose to keep it root-attached as shown in (d).

(c) (d)

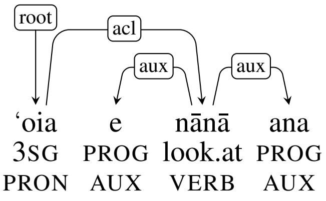


Parses for “to be drunk" using (c) separate *‘ia* for both modern and historical representations and (d) fossilized passive suffixes (e.g. -*mia*).

4.3.3 Dependent Clauses

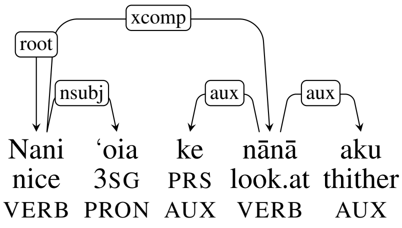
There are numerous issues involved in tagging dependent clauses. For example, Hawaiian has sentences similar to English *tough*-constructions like *Linguists are tough to please* where “an apparently “missing" object" of an embedded infinitival clause is “obligatorily interpreted as coreferential with the matrix subject" (Hicks, 2009, 535). For these, it is unclear whether a dependent should be connected to either a nominal or verbal element. In (e) below, the relative clause *e nānā ana* “(that) t_i is looking" describes the 3rd-person singular pronoun *‘oia*.

(e)



“S/he, who is looking"

(f)



“S/he is nice to look at"

The dependent clause in (e) syntactically parallels that in (f), but with a different semantic interpretation. In (f), the dependent clause describes *‘oia*, similar to the meaning of the English *tough*-construction: “to look at" is a modifier of how “nice" to look at the person is, here expressed by the matrix stative verb *nani*.

5 Conclusions and Future Work

In this paper, we presented a treebank for the Eastern Polynesian language Hawaiian, and used this dataset, along with a treebank in Cook Islands Māori, to construct a crosslingual model to parse both languages. The crosslingual model produced a statistically significant increase in performance for Hawaiian in comparison to the monolingual model. These gains in Hawaiian neither helped nor hurt the performance on CIM.

The zero-shot approach using an unrelated language (English) did not result in any remarkable increase to model performance. A zero-shot model of a related language might still be required to get initial parses when constructing new treebanks from scratch.

Much future work remains for this project. As for the models, we need to perform fine-tuning tests on LLMs (e.g. BERT) and test if they will provide improvements in performance over UDPipe2. Labeling using LLM prompting is more complex, as it touches upon issues of data sovereignty, and the transmission and potential use of this information by the companies that host the LLMs. As for the treebanks themselves, both of them need to be tagged for their morphological features (UFeats), and expanded so that they can dependably label larger collections of data. The Hawaiian language has thousands of pages of text, especially in historical newspaper collections (Shillingford and Parker Jones, 2018), available for parsing; similarly, there is a wealth of legacy information available for Cook Islands Māori that could be analyzed with these parsers.

Evidently the Hawaiian treebank is still extremely small, and the models here will be used for bootstrapping and expanding the treebank. In the work with CIM, we used the Karnes et al. (2023) model to obtain preliminary parses, correct them, and include these newly parsed sentences in the treebank. This approach has been very fruitful in expanding the CIM dataset, from 126 sentences in Karnes et al. (2023) to the present 663. We intend to use this approach for Hawaiian as well, and use these new crosslingual models to accelerate the construction of the Hawaiian treebank even further. Going forth, we intend to coordinate the efforts of the Hawaiian and CIM teams during the annotation, so that any corrections due to improved understanding of linguistic structures can find their way into both sets.

We also intend to expand the domains from which sentences come from to include more varieties of data (e.g. spoken sentences typical of transcribed interviews), so that we can ultimately release these models to the interested stakeholders in their respective communities.

At present, our priority for the Hawaiian parser is to facilitate its access and use by community scholars and organizations involved in language revitalization and pedagogical work. Heeding present intra-community concerns about data stewardship (Alegado et al., 2023), we hope to arrive at a wider consensus among stakeholders before releasing the Hawaiian model for wider distribution and use. As for the CIM, sharing the treebank and the model publicly presents similar concerns, and more consensus is needed before its release. This work is part of a larger project to train linguists and NLP specialists in the Cook Islands, who can collaborate with other scholars from Polynesia in the documentation of their languages.

We hope that this work will be used not only to tag collections in Hawaiian and CIM, but also to foster work in NLP in other Polynesian and Indigenous languages, accelerating documentation work to contribute to language revitalization, normalization, and reclamation efforts in the Pacific and worldwide.

Limitations

The treebanks are largely restricted to written data. While some sentences come from oral interviews, the parsers may still face issues parsing unmodified depictions of spoken language. This limits their applications to naturalistic speech data.

Replicating this project might be difficult in some communities given computational resource demands. To calculate our results, we required 207 hours of GPU time (NVIDIA Tesla K80).

Acknowledgments

We would like to express our heartfelt thanks to all those involved in Hawaiian revitalization, without whose labor this work would not be possible. Mahalo nui loa.

References

Rosie Alegado, Katy Hintzen, Sara Kahanamoku, and Kaleo Hurley. 2023. Kūlana noiʻi: Indigenous data stewardship in hawaiʻi.

Kaavya Chaparala, Guido Zarrella, Bruce Torres Fischer, Larry Kimura, and ʻŌiwi Parker Jones. 2024. Mai Hoʻomāuna i ka ʻAi: Language Models Improve Automatic Speech Recognition in Hawaiian. *arXiv preprint arXiv:2404.03073*.

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.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Kokaʻua, Syed Tanveer, and Isaac Feldman. 2022a. Development of automatic speech recognition for the documentation of Cook Islands Māori.

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano. 2022b. Managing data workflows for untrained forced alignment: examples from costa rica, mexico, the cook islands, and vanuatu. *The Open Handbook of Linguistic Data Management*, 35.

Rolando Coto-Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33, Dunedin, New Zealand.

Cheikh M Bamba Dione. 2021. Multilingual Dependency Parsing for Low-Resource African Languages: Case Studies on Bambara, Wolof, and Yoruba. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92.

Samuel H. Elbert and Mary Kawena Pukui. 1979. *Hawaiian Grammar*. Univ. of Hawaii Press.

Aoife Finn, Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2022. Developing a part-of-speech tagger for te reo māori. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 93–98.

GeminiTeam, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gabriel H. Gilbert. 2023. Directionals in Spoken Hawaiian: A Corpus Analysis. B.A. thesis, Dartmouth College.

Glyn Hicks. 2009. Tough-constructions and their derivation. *Linguistic Inquiry*, 40(4):535–566.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Kelsea Kanohokuahiwi Hosoda. 2019. *Hawaiian morphemes: Identification, usage, and application in information retrieval*. Ph.D. thesis, University of Hawaiʻi at Manoa.

Jesin James, Rolando Coto-Solano, Sally Akevai Nicholas, Joshua Zhu, Bovey Yu, Fuki Babasaki, Jenny Tyler Wang, and Nicholas Derby. 2024. Development of community-oriented text-to-speech models for māori ʻavaiki nui (cook islands māori). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4820–4831.

Peter-Lucas Jones, Keoni Mahelona, Suzanne Duncan, and Gianna Leoni. 2023. Kia tangata whenua: Artificial intelligence that grows from the land and people.

Kauanoe Kamanā and William H. Wilson. 2001. "Mai Loko Mai O Ka ʻIʻini: Proceeding from a Dream": The ʻAha Pūnana Leo Connection in Hawaiian Language Revitalization. In *The Green Book of Language Revitalization in Practice*, pages 147–176. Brill.

Clinton Kanahele. 1970. Clinton Kanahele Collection. University Archives, Joseph F. Smith Library.

Sarah Karnes, Rolando Coto-Solano, and Sally Akevai Nicholas. 2023. Towards universal dependencies in cook islands māori. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 124–129.

Tūreiti Keith. 2024. Work in progress: Text-to-speech on edge devices for te reo māori and ʻōlelo hawaiʻi. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, pages 421–426.

Larry L. Kimura. 1983. Native Hawaiian Culture. In *Report on the Culture, Needs and Concerns of Native Hawaiians*, volume 1 of *Native Hawaiians Study Commission*. Department of the Interior.

R. Keao NeSmith. 2019. Take My Word: Mahalo No i Toʻu Matua Tane. *Linguapax Review*, 7:93–111.

Sally Akevai Nicholas. 2017. *Ko Te Karāma o Te Reo Māori o Te Pae Tonga o Te Kuki Airani: A Grammar of Southern Cook Islands Måori*. Ph.D. thesis, University of Auckland.

Sally Akevai Nicholas and Rolando Coto-Solano. 2019. Glottal variation, teacher training and language revitalization in the cook islands. In *Proceedings of the 19th International Congress of Phonetic Sciences, University of Melbourne, Australia*, pages 3602–3606.

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. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 4034–4043.

OpenAI. 2022. Introducing ChatGPT.

Yuko Otsuka. 2005. History of polynesian languages. *Linguistics*, 345:267–296.

Mary Kawena Pukui and Samuel H. Elbert. 1986. *Hawaiian Dictionary: Hawaiian-English, English-Hawaiian*. University of Hawaii Press.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Neha Ramsurrun, Rolando Coto-Solano, and Michael Gonzalez. 2024. Parsing for mauritian creole using universal dependencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12622–12632.

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F Gerardi. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58.

Brendan Shillingford and ʻŌiwi Parker Jones. 2018. Recovering missing characters in old Hawaiian writing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4929–4934.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2025. Universal Dependencies 2.15 Models.

Tai Tepuaotera Turepu Carpenter and Clive Beaumont. 1995. *Kai Kōrero*.

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.

William H Wilson. 1981. Developing a standardized Hawaiian orthography. *Pacific Studies*, 4:19–19.