Development of Natural Language Processing Tools for Cook Islands Māori

Rolando Coto-Solano[1], Sally Akevai Nicholas[2], and Samantha Wray[3]

[1]School of Linguistics and Applied Language Studies
Victoria University of Wellington
Te Whare Wānanga o te Ūpoko o te Ika a Māui
rolando.coto@vuw.ac.nz
[2]School of Language and Culture
Auckland University of Technology
sally.nicholas@aut.ac.nz
[3]Neuroscience of Language Lab
New York University Abu Dhabi, United Arab Emirates
samantha.wray@nyu.edu

Abstract

This paper presents three ongoing projects for NLP in Cook Islands Māori: Untrained Forced Alignment (approx. 9% error when detecting the center of words), automatic speech recognition (37% WER in the best trained models) and automatic part-of-speech tagging (92% accuracy for the best performing model). These new resources fill existing gaps in NLP for the language, including gold standard POS-tagged written corpora, transcribed speech corpora, and time-aligned corpora down to the phoneme level. These are part of efforts to accelerate the documentation of Cook Islands Māori and to increase its vitality amongst its users.

1 Introduction

Cook Islands Māori has been moderately well documented with two dictionaries (Buse et al., 1996; Savage, 1962), a comprehensive description (Nicholas, 2017), and a corpus of audiovisual materials (Nicholas, 2012). However these materials are not yet sufficiently organized or annotated so as to be machine readable and thus maximally useful for both scholarship and revitalization projects. The human resources needed to achieve the desired level of annotation are not available, which has encouraged us to take advantage of NLP methods to accelerate documentation and research.

1.1 Minority Languages and NLP

Lack of resources makes it difficult to train data-driven NLP tools for smaller languages. This is compounded by the difficulty in generating input for Indigenous and endangered languages, where dwindling numbers of speakers, non-standardized writing systems and lack of resources to train specialist transcribers and analysts create a vicious cycle that makes it even more difficult to take advantage of NLP solutions. Amongst the hundreds of languages of the Americas, for example, very few have large spoken and written corpora (e.g. Zapotec from Mexico, Guaraní from Paraguay and Quechua from Bolivia and Perú), some have spoken and written corpora, and only a handful possess more sophisticated tools like spell-checkers and machine translation (Mager et al., 2018).

Overcoming these limitations is an important part of accelerating language documentation. Creating NLP resources also enhances the profile of endangered languages, creating a symbolic impact to generate positive associations towards the language and attract new learners, particularly young members of the community who might not otherwise see their language in a digital environment (Aguilar Gil, 2014; Kornai, 2013).

As for Polynesian languages, te reo Māori, the Indigenous language spoken in Aotearoa New Zealand, is the one that has received the most attention from the NLP community. It has multiple corpora and Google provides machine-translations for it as part of *Google Translate*, and tools such as speech-to-text, text-to-speech and parsing are

Rolando Coto Solano, Sally Akevai Nicholas and Samantha Wray. 2018. Development of Natural Language Processing Tools for Cook Islands Māori. In *Proceedings of Australasian Language Technology Association Workshop*, pages 26–33.

under development (Bagnall et al., 2017). Other languages in the family have also received some attention. For example, Johnson et al., (2018) have worked on forced alignment for Tongan.

1.2 Cook Islands Māori

Southern Cook Islands Māori[1] (ISO 639-3 `rar`, or glottology `raro1241`) is an endangered East Polynesian language indigenous to the realm of New Zealand. It originates from the southern Cook Islands (see figure 1). Today however, most of its speakers reside in diaspora populations in Aotearoa New Zealand and Australia (Nicholas, 2018). The languages most closely related to Cook Islands Māori are the languages of Rakahanga/Manihiki (`rkh`, `raka1237`) and Penrhyn (`pnh`, `penr1237`) which originate from the northern Cook Islands. Te reo Māori (`mri`, `maor1246`) and Tahitian (`tah`, `tahi1242`), both East Polynesian languages, are also closely related. There is some degree of mutual intelligibility between these languages but they are generally considered to be separate languages by linguists and community members alike.

Figure 1: Cook Islands (CartoGIS Services et al., 2017)

Cook Islands Māori is severely endangered (Nicholas, 2018, 46). Overall its vitality is between a 7 (shifting) and an 8a (moribund) on the Expanded Graded Intergenerational Disruption Scale (Lewis and Simons, 2010, 2). Among the diaspora population and that on the island of Rarotonga there has been a shift to English and there is very little intergenerational transmission of Cook Islands Māori. The only contexts where the vitality is strong is within the small populations of the remaining islands of the southern Cook Islands, where Cook Islands Māori still serves as the lingua franca (see Nicholas (2018) for a full discussion).

1.3 Grammatical Structure

The following is a selection of grammatical features of Cook Islands Māori as described in Nicholas (2017). Cook Islands Māori is of the isolating type with very few productive morphological processes. There is widespread polysemy, particularly within the grammatical particles. For example, in the following sentence, glossed using the Leipzig Glossing Rules (Bickel et al., 2008), there are four homophones of the particle *i*: the past tense marker, the cause preposition, the locative preposition and the temporal locative preposition.

(1) *I* *mate* *a* *Mere* *i* *te* *mangō*
`PST` be-dead `DET` Mere `CAUSE` the shark
i *roto* *i* *te* *moana* *i*
`LOC` inside `LOC` the ocean `LOC.TIME`
te *Tapati.*
the Sunday

'Mere was killed by the shark in the ocean on Sunday.'

Furthermore, nearly every lexical item that can occur in a verb phrase can also occur in a noun phrase without any overt derivation, making tasks like POS tagging more difficult (see section 2.3). The unmarked constituent order is predicate initial. There are verbal and non-verbal predicate types. In sentences with verbal predicates the unmarked order is VSO. The phoneme paradigm is small, as is typical for Polynesian languages, with 9 consonants and 5 vowels which have a phonemic length distinction.

[1] Southern Cook Islands Māori (henceforth Cook Islands Māori) has historically been called Rarotongan by non-Indigenous scholars. However, this name is disliked by the speech community and should not be used to refer to Southern Cook Islands Māori but rather to the specific variety originating from the island of Rarotonga (Nicholas, 2018:36).

2 CIM NLP Projects Under Development

There is a need to accelerate the documentation of the endangered Cook Islands Māori language, so that it can be revitalized in Rarotonga and Aotearoa New Zealand and its usage domains can be expanded in the islands where it is still the lingua franca. We have begun working on this through three projects: (i) We have used untrained forced speech alignment to generate correspondences between transcriptions and their audio files, with the purpose of improving phonetic and phonological documentation. (ii) We are training speech-to-text models to automatize the transcription of both legacy material and recordings made during linguistic fieldwork. (iii) We are developing an interface-accompanied part-of-speech tagger as a first step towards full automatic parsing of the language. The following subsections provide details about the current state of each project.

2.1 Untrained Forced Alignment

Forced alignment is a technique that matches the sound wave with its transcription, down to the word and phoneme levels (Wightman and Talkin, 1997). This makes the work of generating alignment grids up to 30 times faster than manual processing (Labov et al., 2013). In theory, forced alignment needs a specific language model to function (e.g. an English language model aligning English text). However, untrained forced alignment, where for example Cook Islands Māori audio recordings and transcriptions are processed using an English language model, has been proven to be useful for the alignment of text and audio in Indigenous and under-resourced languages (DiCanio et al., 2013).

In order to use this technique, a dictionary of Cook Islands Māori to English *Arpabet* was built, so that the Cook Islands Māori words could be introduced as new English words into the the alignment tool. Some examples are shown in table 1. The phones of Cook Islands Māori words were matched with the closest English phone in the Arpabet system (e.g. the `T` in *kite* 'to know'). Some Cook Islands Māori phones were not available in English; long Cook Islands Māori vowels were replaced by the equivalent accented vowel in English, and the glottal stop was replaced by the Arpabet phone `T`, as in *ngā'i* 'place'.

The English language acoustic model from FAVE-align (Rosenfelder et al., 2014) was used

CIM	Arpabet
kite	K IY1 T EH1
ngā'i	NG AE1 T IY1

Table 1: Arpabet conversion of CIM data

to align previously transcribed recordings of Cook Islands Māori speech. Figure 2 shows the output of this process, a time-aligned transcription of the audio in the Praat (Boersma et al., 2002) TextGrid format.

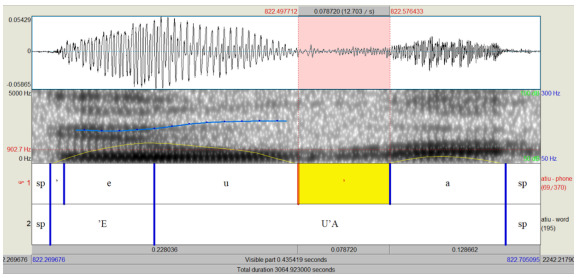

Figure 2: Praat TextGrid for CIM forced aligned text

After the automatic TextGrids were generated, 1045 phonemic tokens (628 vowels, 298 consonants, 119 glottal stops) and the words containing them were hand-corrected to verify the accuracy of the automatic system. Table 2 shows a summary of the results. The alignment showed error rates of 9% for the center of words and 20% for the center of vowels[2]. This error is higher than that observed for other instances of untrained forced alignment (2%, 7% and 3% than that observed for the Central American languages Bribri, Cabécar and Malecu respectively (Coto-Solano and Solorzano, 2016; Coto-Solano and Solórzano, 2017)), but it provides significant improvements in speed over hand alignment.

Type of interval	Error (relative to the duration of the interval)
Words	9% ± 12%
Vowels	20% ± 25%

Table 2: Errors for Untrained Forced Alignment

Utilization of forced alignment as a method for documentation and phonetic research has already

[2]We are currently documenting phonetic variation in vowels using data that was first force aligned and then manually corrected. Because of this, we don't know at this point whether the 20% error is due to phonetic differences between English and CIM vowels, or if it's due to the data itself.

been demonstrated for Austronesian languages. In Nicholas and Coto-Solano (2018), Cook Islands Māori phonemic tokens including vowels, consonants and glottal stops were forced-aligned and manually corrected to study the glottal stop phoneme. This work was able to show that in islands such as ʻAtiu, whose dialect is reported as having lost the glottal stop, the phoneme survives as shorter stop or as creaky voice. The corrected Praat TextGrids are publicly available at Paradisec (Thieberger and Barwick, 2012), in the collection of Nicholas (2012).

2.2 Automatic Speech Recognition

We have begun the training of an Automatic Speech Recognition (ASR) system using the Kaldi system (Povey et al., 2011), both independently and through the ELPIS pipeline developed by CoEDL (Foley et al., 2018). While this work is still in progress, our preliminary results point to the need of custom models for each speaker. As can be seen in table 3, our recordings produced models with very different per-speaker word-error rates. (The "all speakers" model has cross-speaker test set). This, in addition to the paucity of data (approx. 80 minutes of speech for all speakers) makes the task extremely difficult.

Speaker	WER
Female, middle-aged, controlled environment	37%
Female, middle-aged, open environment	55%
Female, elderly, open environment	62%
Male, elderly, open environment	68%
All speakers	64%

Table 3: Errors for Untrained Forced Alignment (per-speaker)

The recording with the best performance was recorded in a very controlled environment (a silent room with a TASCAM DR-1 recorder). The worst recordings were those of elderly speakers who were speaking in their living rooms with open windows. The main issue here is that it is precisely these kinds of recordings (open environments with elderly practitioners of traditional knowledge or tellers of traditional stories) that are of most interest to linguists and practitioners of language revitalization, and it is in those environments where we can see our worst performance. More work on this area is needed (e.g. crowdsourcing the recording of fixed phrases from numerous speakers for more reliable training).

2.3 Part-of-Speech Tagging

We have begun developing automatic part-of-speech (POS) tagging to aid in linguistic research of the syntax of Cook Islands Māori, and to build towards a full parser of the language. To begin our work we hand-tagged 418 sentences (2916 words) using the part of speech tags from Nicholas (2017). This is the only POS annotated corpus of Cook Islands Māori existing to date. The corpus is currently being prepared for public release.

The corpus is currently annotated using two levels of tagging: a more shallow/broad level with 23 tags, and a second, narrower level with 70 tags. For example, the shallow level contains tags like `v` for verbs, `n` for nouns and `tam` for tense-aspect-mood particles. The narrower level provides further detail for each tag. For example, it separates the verbs into `v:vt` for transitive verbs, `v:vi` for intransitives, `v:vstat` for stative verbs, and `v:vpass` for passives.

Classification experiments were carried out in the WEKA environment for machine learning (Hall et al., 2009). We tested algorithms that we believed would cope with the sparseness of the data given the size of the corpus. These algorithms were: *D. Tree (J48)*: an open-source Java-based extension of the C4.5 decision tree algorithm (Quinlan, 2014) and *R. Forest*: merger of random decision trees (with 100 iterations). Included as a reference baseline is a *zeroR* classification algorithm predicting the majority POS class for all words. All algorithms were evaluated by splitting the entire corpus of 2916 words into a 90% set of sentences for training and 10% set for testing.

The model used position-dependent word context features for classification of each word's POS. These included:

- the word (*w*)
- the previous word (*w-1*)
- the word before the previous word (*w-2*)
- the word two before the previous word (*w-3*)
- the following word (*w+1*)

Model	Accuracy
Shallow/broad POS tags	
D Tree (j48)	87.59%
R Forest (I = 100)	**92.41%**
zeroR (baseline)	21.03%
Narrow POS tags	
D Tree (j48)	80.00%
R Forest (I = 100)	82.41%
zeroR (baseline)	15.52%

Table 4: Accuracy of POS tagging models. Performance is reported in accuracy (per-token)

A comparison of the classification algorithms using the features above is shown in Table 4. As seen here, the current top performing classifier that we have identified is a *Random Forest* classifier. This algorithm performs best when the POS tags are less informative; that is, it performs best on the shallow/broad tags with an accuracy of 92.41%. Despite the fact that the narrow tags do not collapse across types and therefore are more difficult to classify, the best performer for the narrow tags is also the *Random Forest* classifier. This performance is comparable to other POS tagging tasks for under-resourced languages for which a new minimal dataset was manually tagged as the sole input for training, such as 71-78% for Kinyarwanda and Malagasy using a Hidden Markov Model (Garrette and Baldridge, 2013). It is also comparable to POS tagging for related languages with relatively larger corpora, such as Indonesian (94% accuracy, with 355,000 annotated tokens) (Fu et al., 2018).

To obtain an assessment of directions for future work aimed at improving the model, we also determined the most commonly confused tags by consulting a confusion matrix. The most common errors for the top performer (Shallow/broad tags as classified by the Random Forest) are seen in table 5. Recall from section 1.3 that grammatically, lexical items which occur in a noun phrase can also occur in a verb phrase with no overt derivational marking. This explains the fact that the majority of errors occurred as the result of confusion between `v` and `n`.

After training the model, we built a JavaServer Pages (JSP) interface to demo the model and obtain POS tagged versions of new, raw untagged sentences. This is illustrated in figure 3 below. The interface is in the process of being prepared for public launch.

Error type assigned tag ⇒ correct tag	% of errors
n (NOUN) ⇒ v (VERB)	23%
tam (tense aspect mood) ⇒ prep	9%
prep ⇒ tam (tense aspect mood)	5%
v (VERB) ⇒ n (NOUN)	5%

Table 5: Most common POS tagger errors for shallow/broad tags for top-performing tagger (Random Forest)

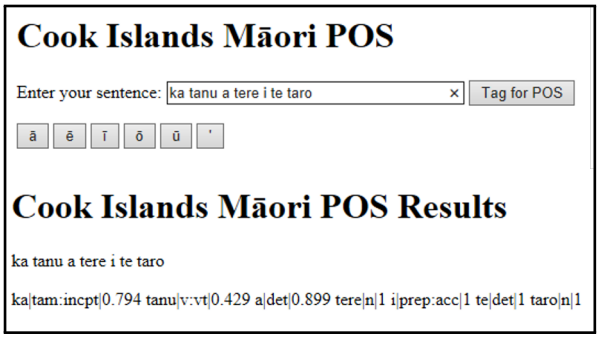

Figure 3: Interface for the POS tagger

The assignation of parts of speech is a very difficult task in Cook Islands Māori not only because of its data-driven nature, but because the orthography of Cook Islands Māori has not been fixed until recently. As detailed in Nicholas (2013), historically and even today there are numerous variations in spelling. The glottal stop is frequently omitted in spelling, as are the macrons for vocalic length. As a result, words like *'e* 'nominal predicate marker' can be written as *'ē* or *e*, increasing the homography of the language. Figure 4 shows the JSP interface attempting to tag two variants of the greeting *'Aere mai* 'Welcome'. The first is spelled according to the current orthographic guidelines, with a glottal stop character at the beginning, and the first word gets tagged correctly as an intransitive verb. The second word, however, is spelled without the glottal stop, which leads the model to misidentify it as a noun. Future work includes experiments on how to best tackle this problem by utilizing naturally occurring variety written as spontaneous orthographies, looking at solutions already found for languages with non-standardized writing systems such as colloquial Arabic (Wray et al., 2015).

Figure 4: Error when tagging words without a glottal stop

2.4 Language Revitalization Applications

In addition to the scholarly advances these projects will support, they also have utility for the revitalization of Cook Islands Māori. The forced alignment work will support the teaching of the phonetics, phonology, and 'pronunciation', as well as improve community understanding of variation. The consumption of audio visual material with closed captions in the target language is known to be beneficial for language learning (Vanderplank, 2016), and automatic speech recognition will greatly increase the quantity of such material available to learners. The increased size of the searchable corpus of Cook Islands Māori will facilitate the production of corpus-based and multimedia pedagogical materials. Similarly, the POS tagged data will provide further benefits for the accurate design of pedagogical materials and the linguistic training of teachers who speak Cook Islands Māori. Additional tools that can be developed using this well annotated corpus include text-to-speech technology and chatbot applications.

3 Future work

There is much work to be done to move forward in these three projects. The next step for the POS tagging is to add resilience to cope with the non-stardardized writing it will most commonly find. As for the speech recognition, crowdsourcing similar to that carried out by Bagnall et al (2017) might help increase the amount of controlled, pre-transcribed audio available for speech recognition training. Furthermore, we are investigating the incorporation of algorithms for treatment of audio data with low signal-to-noise ratio to improve audio quality which theoretically should lower the high word error rate for recordings conducted in an open, noisy environment.

4 Conclusions

This paper summarizes ongoing work in the application of NLP techniques to Cook Islands Māori, including untrained forced alignment for producing time-aligned transcriptions down to the phoneme, automatic speech recognition for the production of transcriptions from recordings of speech, and part-of-speech tagging for producing tagged text. We have given an overview of the state of these projects, and presented ideas for future work in this area. We believe that this interdisciplinary work can accelerate and enhance not only the documentation of the language, but can ultimately bring more of the Cook Islands community in contact with its language and help in its revitalization.

Acknowledgments

We wish to thank Dr. Ben Foley and Dr. Miriam Meyerhoff of CoEDL for their support in numerous aspects of this project, as well as three anonymous reviewers for their helpful comments. We also wish to thank Jean Mason, Director of the Rarotonga Library, Teata Ateriano, principal of the School of Ma'uke, and Dr. Tyler Peterson from Arizona State University for their continued support in the documentation of Cook Islands Māori.

References

Yāsnaya Elena Aguilar Gil. 2014. para qu publicar libros en lenguas indgenas si nadie los lee? e'px. http://archivo.estepais.com/site/2014/para-quepublicar-libros-en-lenguas-indigenas-si-nadie-los-lee/.

Douglas Bagnall, Keoni Mahelona, and Peter-Lucas Jones. 2017. Kōrero Māori: a serious attempt at speech recognition for te reo Māori. Paper presented at the Conference 2017 NZ LingSoc, Auckland.

Balthasar Bickel, Bernard Comrie, and Martin Haspelmath. 2008. The leipzig glossing rules. conventions for interlinear morpheme by morpheme glosses. *Revised version of February* .

Paul Boersma et al. 2002. Praat, a system for doing phonetics by computer. *Glot international* 5.

Jasper E Buse, Raututi Taringa, Bruce Biggs, and Rangi Moeka'a. 1996. *Cook Islands Māori dictionary with English-Cook Island Māori finderlist*. Pacific Linguistics, Research School of Pacific and Asian Studies, Australian National University, Canberra, ACT.

CartoGIS Services, College of Asia and the Pacific, and The Australian National University. 2017. Cook islands. http://asiapacific.anu.edu.au/mapsonline/base-maps/cook-islands [Accessed 2017-10-06].

Rolando Coto-Solano and Sofía Flores Solórzano. 2017. Comparison of two forced alignment systems for aligning bribri speech. *CLEI Electron. J.* 20(1):2–1.

Rolando Coto-Solano and Sofia Flores Solorzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Káñina* 40(4):175–199.

Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134(3):2235–2246.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Sihui Fu, Nankai Lin, Gangqin Zhu, and Shengyi Jiang. 2018. Towards indonesian part-of-speech tagging: Corpus and models. In *LREC*.

Dan Garrette and Jason Baldridge. 2013. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL 2013*. Atlanta, USA, pages 138–147.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. 2009. The weka data mining software: an update. *ACM SIGKDD explorations newsletter* 11(1):10–18.

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation & Conservation* .

András Kornai. 2013. Digital language death. *PloS one* 8(10):e77056.

William Labov, Ingrid Rosenfelder, and Josef Fruehwald. 2013. One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language* 89(1):30–65.

M. Paul Lewis and Gary F. Simons. 2010. Making EGIDS assessment for the ethnologue. www.icc.org.kh/download/Making_EGIDS-_Assessments_English.pdf [Accessed 2017-07-20].

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018. Challenges of language technologies for the indigenous languages of the americas. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)* .

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

Sally Akevai Nicholas. 2012. Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects) (sn1). Digital collection managed by PARADISEC. [Open Access] DOI: 10.4225/72/56E9793466307 http://catalog.paradisec.org.au/collections/SN1.

Sally Akevai 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* 15:36–64.

Sally Akevai Nicholas and Rolando Coto-Solano. 2018. Using untrained forced alignment to study variation of glottalization in cook islands mori. In *NWAV-AP5*. University of Brisbane.

Sally Akevai Te Namu Nicholas. 2013. Orthographic reform in cook islands māori: Human considerations and language revitalisation implications. *3rd International Conference on Language Documentation and Conservation (ICLDC)* .

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, EPFL-CONF-192584.

J Ross Quinlan. 2014. *C4. 5: programs for machine learning*. Elsevier.

Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. FAVE (Forced Alignment and Vowel Extraction) Program Suite.

Stephen Savage. 1962. *A dictionary of the Maori language of Rarotonga / manuscript by Stephen Savage*. New Zealand Department of Island Territories, Wellington.

Nicholas Thieberger and Linda Barwick. 2012. Keeping records of language diversity in melanesia, the pacific and regional archive for digital sources in endangered cultures (paradisec). *Melanesian languages on the edge of Asia: Challenges for the 21st Century* pages 239–53.

Robert Vanderplank. 2016. Effects of and effects with captions: How exactly does watching a tv programme with same-language subtitles make a difference to language learners?. *Language Teaching* 49(2):235.

Colin W Wightman and David T Talkin. 1997. The aligner: Text-to-speech alignment using markov models. In *Progress in speech synthesis*, Springer, pages 313–323.

Samantha Wray, Hamdy Mubarak, and Ahmed Ali. 2015. Best Practices for Crowdsourcing Dialectal Arabic Speech Transcription. In *Proceedings of Workshop on Arabic Natural Language Processing*.