



Māori Pronunciation Dictionary Building Tool

Developers
James Coppard, Brendon Joe

Supervisors
Catherine Watson, Peter Keegan

Introduction

Te Reo Māori is the language of Māori, the indigenous people of New Zealand. Māori is considered an **under-resourced** language in the context of speech technologies.

An integral component of speech technology for a language is a **pronunciation dictionary (PD)** which maps words to their pronunciations.

Previous work has been done on producing CLI tools for Māori pronunciation generation, which we have extended into a **web-based tool**.

Māori Linguistics

Māori was originally only a spoken language and had no writing system. The Latin alphabet was introduced when British missionaries arrived in the 1800s.

Syllables - of the form: (C)V(V), e.g. a, ae, ke, wha, mā
All Māori syllables are **open** (end with a vowel)

Consonants (C) h k m n p r t w ng wh

Vowels (V) a e i o u

Long Vowels ā ē ī ō ū

Diphthongs - a combination of two non-identical vowels into a single sound: ae, ai, ao, au, ei, oi, oe, ou

Stress - the most prominent syllable(s) we hear in a word
Syllable stress can be **applied** using **Biggs' rules [1]**

Stress is applied once in groups of **four vowels**, in the order:

First double vowel → First non final diphthong → First vowel

However, sometimes a Māori word will **not** follow these rules.

Within a dialect, letter-to-phoneme conversions do not vary, so the only variable parts of a pronunciation are the syllable boundaries and stress position.

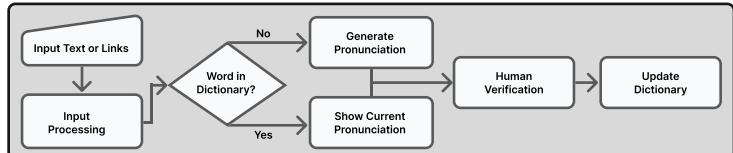
Design

A Māori PD building tool was previously developed using Python, which is run via a command line [2]. We identified a few areas that could be improved:

- Multi-morphemic and compound word stress generation
- Local text file word data storage
- Manual pronunciation editing

User Workflow

We have taken inspiration from a user workflow of a proven pronunciation editor, LEXiTRON Pro [3]:



Database

The previous text-based storage was inflexible and quite hard to work with. We have chosen MongoDB as our database, which will help us adapt to change in the future.

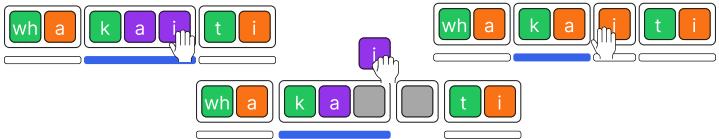
Architecture

Our frontend was written in React, which allowed us to make a complex and beautiful UI. Flask + Python were used in the backend so we could leverage the existing language tooling.

Features

Pronunciation Editor

Our advanced drag and drop based pronunciation editor allows full granular control of syllable boundaries and stress.



Bulk Word Adding and Verification

The user can paste blocks of text or links which get scraped. Text then gets processed into individual words, and non Māori words are removed at a **97% success rate**.

Word	Exists	Verified
maori	No	Yes
home	No	No
note	No	No
kumara	No	No
kupe	No	No

Orthography:
Spelling: kumara
[Edit](#)
Details:
Verified: No
Source: William's Dictionary
Pronunciation:
[Play](#)

Controls:
[Add](#) [Skip](#) [Reject](#)

- User goes through each word one by one.
- Spelling and pronunciation can be edited if they are wrong.
- Each word can be added or rejected

Exporting

Users can export the dictionary to various text formats, namely: JSON, TXT, MaryTTS and Festival

Administrator Control

- Management of user access control at different permission levels
- All additions and edits to words are saved and displayed

Results

User Testing

- Pronunciation editing features of the tool are intuitive to users with some knowledge of Māori linguistics
- Enforced word syllable boundaries and stress on UI help to minimise invalid changes made by users

Pronunciation Generation

- Compared with **957 verified words**, generated word pronunciations have a **95% match rate**
- Verifying **pronunciations generated** by the tool is **23% faster** compared to manual transcription* and checking

*Performed by an expert in Māori linguistics

Conclusions and Future Works

We have succeeded in the future work goals set out by previous research through implementing a **GUI**, individual **pronunciation editing**, and **multi-stress application**. Our simple interface and **centralised server** enable linguists worldwide to contribute to a unified Māori PD. Our **verification tools** can produce valuable pronunciation data, paving the way for future work on reassessing the viability of **machine learning** methods for generating Māori word pronunciations.

References

- [1] Let's Learn Māori (Revised) - B. Biggs
- [2] Generating a Māori Pronunciation Dictionary - R. Berriman
- [3] LEXiTRON-Pro Editor - S. Klaithin et al.