Phonetic Modelling of the Borrowing Process for Māori Loanwords

Prepared by: Liam Morris
Supervised by: Assoc. Prof. Andrew Mason, Dr. Catherine Watson, and Dr. Peter Keegan

THE UNIVERSITY OF AUCKLAND NEW ZEALAND | ENGINEERING

National SCIENCE Challenges

PARININIHI KI WAITOTARA

INTRODUCTION

BACKGROUND

This project is part of the Science for Technological Innovation National Science Challenge.

Missing and incomplete records mean that thousands of Māori descendants who are rightful shareholders in ancestral land can not be tracked down. Parininihi ki Waitotara (PKW) is one such Māori incorporation which has lost track of around half of its 10,300 owners.

Algorithms to search death certificates, newspapers, and Māori Land Court records have been developed to search for missing shareholders. However, many connections are missed during this search due to family members that have both an English name and a Māori Name.

These names are often the result of a borrowing process which modifies English names to follow the structure of the Māori language. In general, this process is called 'borrowing', and borrowed words are called 'loanwords'

Some example borrowing pairs:

Patrick → Patariki

Adam → Arama, Ātama, Atana

These borrowings reveal three things:

1. Māori words always end in a vowel sound.
2. Māori words do not typically have adjacent consonants.
3. There is no one 'correct' way of borrowing a word.

But what about more complicated borrowings?

Fowler → Wharaora

Francis → Whereakiko

AIMS

Begin the development of a software package capable of the following:

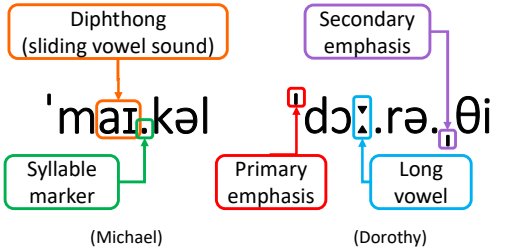
- Applying general rules to match sounds between English and Māori word pairs.
- Analysing known loanwords to determine rules used to modify borrowed English words.
- Discovering unknown rules used during the borrowing process.
- Determining the likelihood of any pair of English and Māori words being a borrowing.

METHODS

IPA TRANSCRIPTION

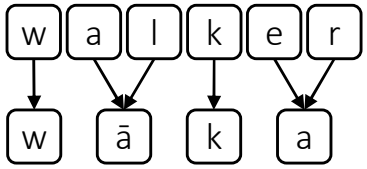
Many English names are not pronounced as they appear. This makes the matching of sounds between borrowing pairs difficult for a computer. The International Phonetic Alphabet (IPA) is used to solve this problem.

IPA is an alphabet where symbols represent sounds used in human speech. IPA was used to transcribe borrowing pairs so that the modelling is sound-based.

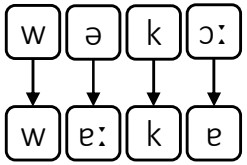


Transcribing words with IPA means borrowings can now be more easily matched based on sound composition rather than spelling.

Mapping with English/Māori:



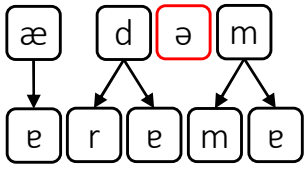
Mapping with IPA:



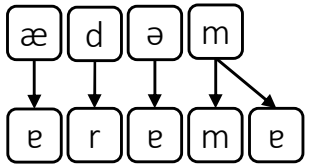
PHASE 1 – RULE DISCOVERY

The first phase aims to learn a set of possible mapping rules from public borrowings databases, including those in Ngā Kupu Arotau - eweri tāima: Loanwords in Māori 1842-1952 by Ka'ai, T. & Moorfield, J.. (2009).

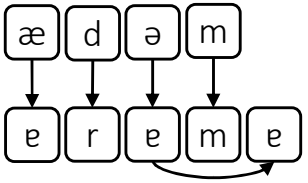
Most borrowings have multiple ways sounds can be matched between the words.



In this instance the schwa (ə) is lost in the borrowing. This is due to the argument that schwas are just filler sounds and aren't actually enounced.

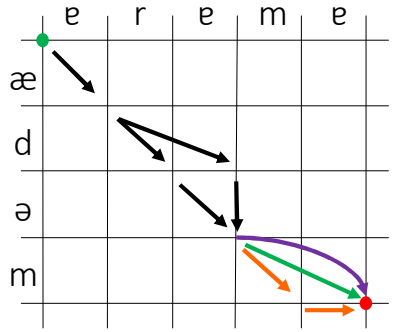


In this case the schwa becomes the second ɐ sound in Arama.



It's also possible that the appended ɐ sound is actually dependent on the previous vowel and not the previous consonant.

These alternate mappings can be described as different paths through a network where each arc represents a different mapping:



The coloured arcs represent three instances of the same result that could be produced via different rules.

These arcs are created by searching for applicable regular expressions (regex) like these:

Appended vowel matches previous vowel:
[bcdf...]$ > (?<=([ɐɛɔʉi]))[fwpt...]\1$

Any vowel can be appended to a consonant:
[bcdf...]$ > [fwpt...] [ɐɛɔʉi]$

Any consonant can become a consonant AND any vowel can be appended:
[bcdf...] > [fwpt...]
> [ɐɛɔʉi]$

By finding all possible paths through the network for every known borrowing pair, the occurrences of different mappings can be counted and used as an estimate of how useful they are. E.g.:

b -> p due to [bcdf...] > [fwpt...] - 59.87
m -> m due to [bcdf...] > [fwpt...] - 76.34
m -> mɐ due to [bcdf...] > [fwpt...][ɐɛɔʉi] - 20.74
...

PHASE 2 - OPTIMISATION

Phase 2 is currently in progress, and aims to maximise the 'fit' to the data while using the least amount of rules possible.

This is currently being done by assigning each mapping:

1. A probability based on its frequency in phase one.
2. A penalty value based on how specific the rule is.

The sum of shortest path lengths for each borrowing pair, balanced against the number of mappings is used as the objective function for gradient descent.

The aim of this is to remove more specific rules when the mappings can be explained just as well by more general ones.

However, generating and solving millions of shortest path problems is proving to be a computationally expensive method so other solutions will be investigated.

CONCLUSIONS

Development of the software package has begun, and the following features have been implemented:

- Automated fetching of IPA transcriptions for English words from the Wiktionary.com API.
- Software for generating and solving shortest path problems from a set of general rules.
- Generated list of possible borrowing mappings.
- Partial implementation of rule optimisation routines.

More work is required to improve the set of borrowing rules used to explain known mappings.