

## Intro

For our project we've decided to build a text reader with a Māori vocabulary training component. It will allow users to open a variety of text file formats (txt, ePub etc) occasionally substituting words with their Māori translation. The app will include a library of public domain eBooks for users to read. We see this as an option to replace the users default text reader, something akin to adobe pdf reader but for all text.

Our app will utilize the Diglot Weave approach, replacing words in the text with words from a language they wish to learn. The Diglot Weave approach aims to help users learn words by placing them in context. At first, we will focus on the Māori language, with the goal of extending to other languages in the future.

## The team

Our team consists of Jet, Jake, Matt & Cody.

Jake is a third year computer science student with interest in low level programming and design. He has some experience in natural language processing. He is well practised in working with teams from his time as a chef, however this is his first time building an app as a team from scratch.

Jet has almost finished his undergrad computer science and is starting honours next year. He has experience with C++ in the context of graphics and computer vision, as well as designs skills and some artistic ability.

Cody is a third year computer science student with an information science minor, including 4 years of experience in IT. This is his third team based software project. He has experience in NLP, doing research over the summer, which should hopefully prove useful for this project.

Matt is a semi professional fry cook with 28 years of life experience, some of which are in programming. He has heard of C++, but isn't sure why they didn't just call it D, or at least ++C so it would run quicker.

## What datasets are we using?

For our data we need a high quality English to Māori Dictionary, and repository of public domain books. The latter is easily decided, as Project Gutenberg offers a fantastic database of books to draw from. The dictionary is maybe not quite as straight forward, as although many exist, they are not always up to date and well digitised. We propose to get around this by drawing from as many datasets as we can and using it to inform our final dataset for translating. Some notable sources of this data are Te Ngutu Kura, the Māori spell checker

## How will we build it?

We have made a rough diagram to show the architecture for the project

As seen in the diagram we are trying to split the application into three main areas:

- Backend API
- Controller
- GUI

### Backend API

This is further split into three different sections:

- User Profile
- Translator
- Substitute

We aim to keep each part of the backend isolated, in that it will send and receive data through the controller abstraction layer

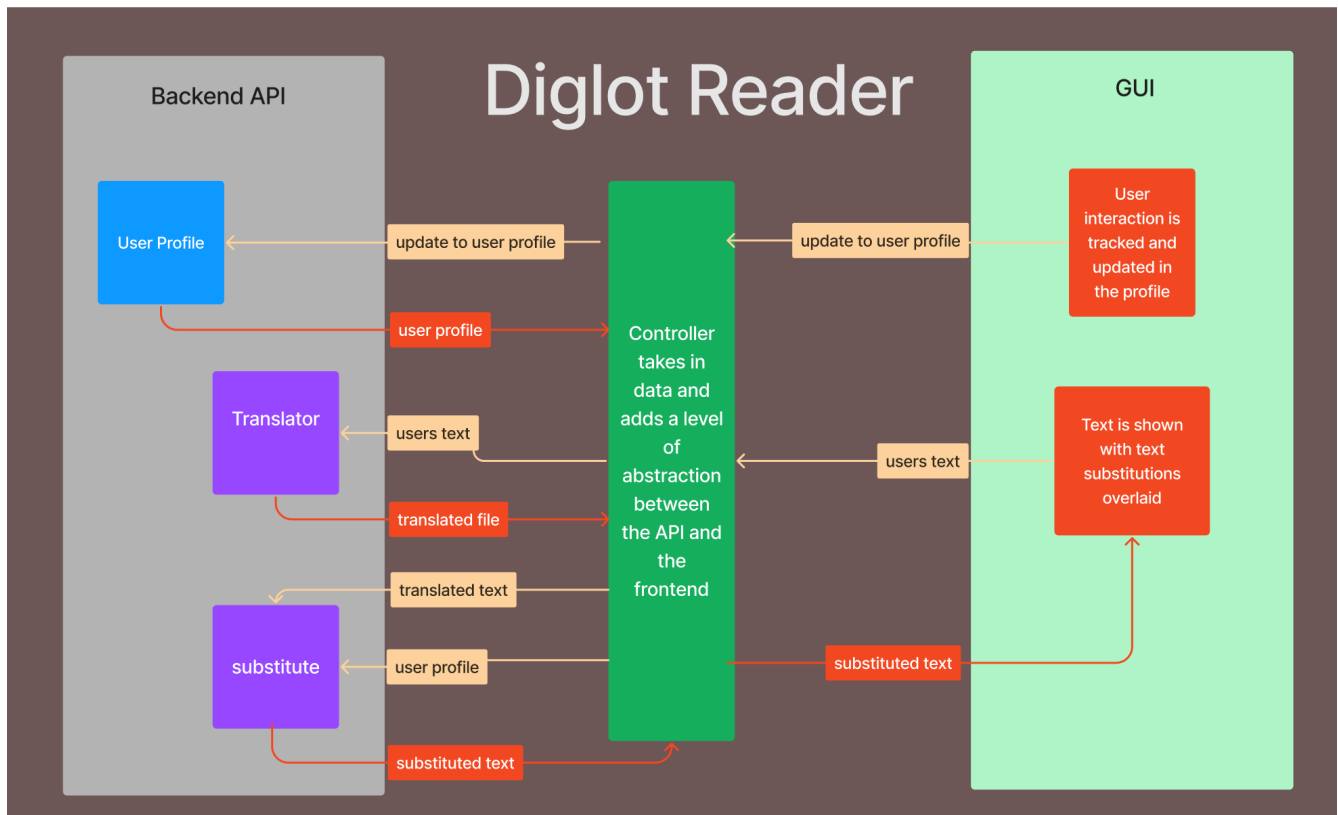


Figure 1: diagram

## Controller

This will handle the data going between different areas of the code. The GUI will send text and any updates to user data to the controller, which will then make the API calls to coordinate the backend. Once the backend has finished its computation, the controller will then send the result to the GUI.

## GUI

From a UI design perspective we are aiming to have a low clutter and simplistic design aimed to be easy on the eyes for long reading sessions. We have made some mockups in figma of some of the features that we envision for the app. These mockups will then be transferred to html and css using the ultralight framework as a html renderer. Figma itself allows for the designs to be exported to html and css, though there will also be some clean up on our end. Ultralight allows us to make an app using a UI framework we understand, html, but be able to easily couple it with C++.



As seen we will have a landing page from which you can open recent books or documents. Additionally there is a sidebar which will show a typical file system to navigate to files, or optionally look through the users library.

## What is our schedule?

Alpha (4 weeks)

Features: \* Ability to open and read a file \* Word substitution with Maori translations \* Display of translations

The Alpha milestone focuses on creating a minimum viable product. It involves three main components: Data organization, Frontend setup for the app’s template, and the implementation of the translation and controller architecture. This phase heavily relies on sequence, with frontend components like the hover feature requiring mappings to show the correct translations on mouseover. Other frontend components such as the reader view and main menu also need to be integrated.

Note: The schedule accounts for the mid-semester break.

Beta (2 weeks)

Features: \* Library tab \* Blacklist / whitelist \* Recent documents on the homepage \* Stats menu for progression tracking \* Word learning tests

The Beta phase aims to expand the app with primary components. Each feature can be implemented in parallel, but there are some dependencies within each feature’s implementation. The focus shifts towards creating a customized learning tool rather than a simple word-switching reader.

Final Product (3 weeks)

Features: \* Dark mode / Light mode support \* Support for .docx and .pdf files \* Legend for currently switched words \* App documentation

The Final Product phase introduces quality of life features. This allows some flexibility in the schedule, enabling us to address any unforeseen challenges encountered during the Beta phase. The tasks in this phase are organized to minimize dependencies and avoid overcommitting resources.

Testing:

Testing is scheduled for the second half of each sprint, ensuring that the team has enough time to address any issues discovered during the testing phase.

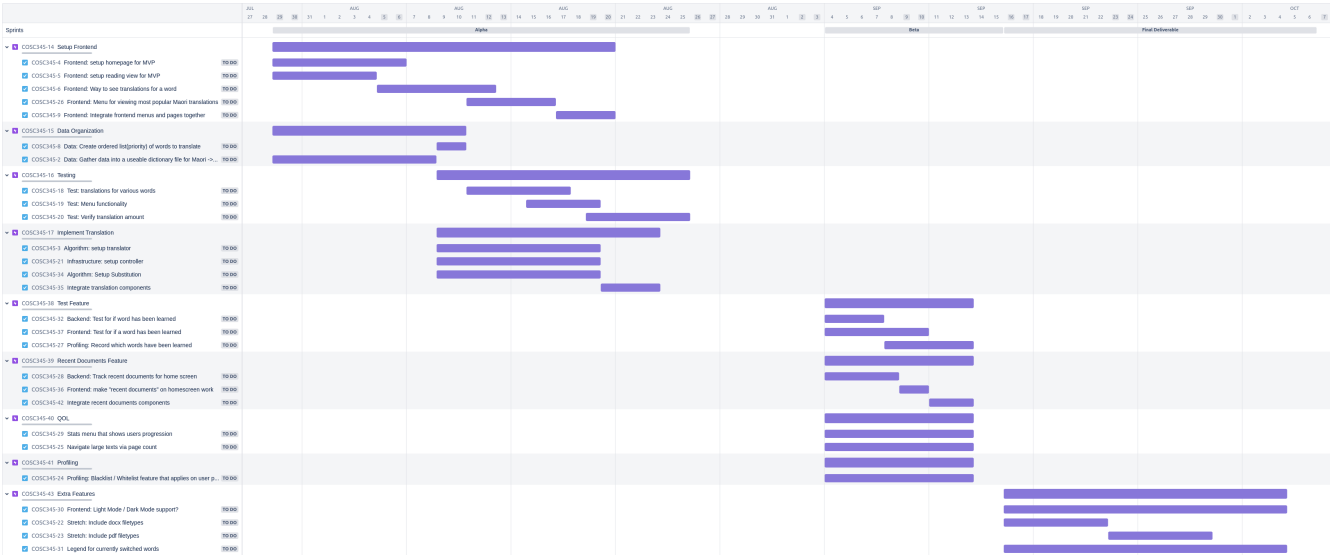


Figure 2: gantt\_chart

## What apps already exist?

We found in our research that this idea has been looked at many times before. The core concept of our app has a basis in research, with the best known promotion of this concept being the Diglot Weave technique as it became to be known, start with R. Burling in his “outlandish” 1968 paper. Other similar techniques have been known as the “Clockwork Orange” technique coming from research in the late 90s that had noticed those who read A Clockwork Orange learned and retained Russian slang words very well.

As for commercial ventures there are some notable examples. Prismatext ([prismatext.com](https://prismatext.com)) cites the Diglot Weave technique as a core component of its technology. It has public domain books that have been edited by translators to slowly transition into a variety of different languages. Although it itself does not offer the ability to open your own text, and it does not appear to offer Māori as an option.

We also found that the Clockwork Orange Technique was being used, rather amusingly, to write children's books by onethirdstories ([onethirdstories.com](https://onethirdstories.com)). They focus largely on offering physical books and only aimed at younger readers. They also have a fairly limited language selection, Māori not being on the list.

We also found Weeve (<https://shop.weeve.ie/>) which offers both an app and physical books for sale. This one is of particular interest as it is an Irish company that offers texts in Irish, which faces similar challenges to adoption as Māori does. Ultimately, they rely heavily on preprepared text to offer their translations, and don't have the ability to open your own.

Other more promising searches from a feature point of view do appear, though often lead nowhere. We have loomVue(<https://loomvue.medium.com/loomvue-sneak-peak-320c6e385dba>) that seemingly offered an extension to do this but it doesn't appear obvious if this is still around or indeed ever made it to market. They did however take a machine translation approach and dealt with alignment issues.

Through all of this, we find a mixture of the features we like, yet nothing that is: free, open source, and offers Māori as an option.

## Is there interest in our app?

Measuring interest is not exactly straight forward. The fact that there are commercial companies based around a similar proposition suggests that there is interest in the core idea of mixing a second language into a first language's texts. Although it is also notable that none of them do this with machine translation on a commercial scale.

We also believe that we offer something unique, that is the ability to read one's own text, and that the language that is on offer is Māori.

There is also a rather different cultural argument to its use in New Zealand, as New Zealand English (NZE) has a strong propensity to use Māori loanwords in place of their English counterparts. Data from the 2021 General Social Survey (GSS) showed that the ability of New Zealanders above the age of 15 to speak some Māori had increased. Between 2018 and 2021 the proportion of people able to speak more than just a few words or phrases in Māori had increased from 24% to 30%.

The same survey showed that 3 in 5 New Zealanders think that Te Reo should be a core subject in primary school. This suggests that there is at least some demand for Te Reo resources