PROJECT PROPOSAL

Extraction of Structured Information from Dictionary PDFs

Caleb Smith
Josephine Paculio
Joshua Dib
Roger Bernardo

Executive Summary

Currently in the industry of linguistic analysis dictionaries must be manually converted into a digital format and placed into a spreadsheet before the advantages of technologies such as formulae and quality of life feature can be utilised to aid researchers in their analysis. This act is a tedious and time-consuming activity which not only requires a strong eye for detail and an advanced level of computer dexterity it most importantly ruins the researches workflow and detracts from meaningful hours that could be spent making progress and dissecting the content being prepared. To serve as a fix for this issue an automatic OCR tool will be developed to aid in the extraction of dictionary entries for Dr Rachel Hendery and her team to use in their works.

Found with this document details will be explained about on three proposed solutions, those being an independent web application that offers a convenient and accessible service, an online focused mobile application featuring a more robust toolset for dictionary and OCR result preservation and leaves room for scalability, and finally a desktop application providing users with a straight forward, local process to carry out the task at hand in the comfort of their own PC.

From these, solution 1 being a simple web application emerges as the optimal recommendation due to the ease of use of has a web-based approach not requiring the movement or installation of files, easier development thanks to the wide availability of feature-rich packages. When compared to a desktop application the platform independence is a must in modern day where operating systems vary from user to user and easier development to ensure it is complete on time. Moreover, it does not clutter itself with additional, unused, unwanted features when compared to more situational approach of the mobile app such as integration with a living database which is not specifically needed for the task and adds towards development time.

Furthermore, the product functionality which is universal across all solutions with their related user actions will be listed as well as a development timeline for the described features will be clearly defined and used for revision and guideline for optimal progression.

The completion of this document marks the formal start of development as solution approaches and functional requirements are agreed upon by both sides involved, eliminating any prior assumptions and setting out a defined plan that will be followed.

Table of Contents

1. Introduction

As technology becomes more widespread and readily available, applications and processes have been put in place to improve quality of life and provide society with increased ease and efficiency for tasks. Linguistic analysis is one field to see the benefits technology can have through use of automation and algorithm application.

This document will provide the reader with an in-depth look into the project at hand and why it is a problem that requires a solution. Moreover, several proposed development plans will be detailed explaining the technologies each and their associated consequences both positive and potentially negative aspects. Furthermore, functionality will be defined for viewing and allow for a mutual understanding and expectation of what to the product will provide.

Following on from this document a better insight will be gained into the proposed solution going forward and agreement will be made on how to proceed with implementation.

2. Client details and Project background

Doctor Rachel Hendery is an Associate professor of Digital Humanities at Western Sydney University. She is a Linguist who works on how new technological development helps in finding new ways to study and research about language contact and change in the Pacific.

Her Doctor of Philosophy is about observing changes in relative clauses constructions cross-linguistically which is a project about historical typology. Moreover, her undergraduate degree was a Bachelor of Arts in Linguistics and German at University of Canterbury in New Zealand. Furthermore, Dr Hendery also got a Masters of Arts in Comparative Linguistics and German Medieval Literature at Johann-Goethe University in Frankfurt, Germany.

She supervises postgraduate projects about digital humanities such as data visualization, mapping, language, virtual reality, simulation and other topics such as typology, historical linguistics and contact linguistics.

She is currently working on her linguistic ARC project that deals with observing the relation of contact of different communities and the change in language or communication that might have been caused by the interaction between two or more communities.

3. Problem Statement

The linguistic ARC project involves working with large amount of PDF files that consist of different foreign language dictionaries. Optical character recognition (OCR) software is being used as a tool to enhance the quality and readability of these dictionaries. After going through the OCR software, the PDF files are being used as reference for searching words, phrases and translations.

This process is being done by using the search functionality that is featured in the Acrobat reader. However, the search functionality in Acrobat reader and other PDF viewers are only designed to work on a single document at a time.

This limitation makes the process longer since a single search must be repeatedly done to all different PDF files.

4. Project Requirements

Minimum Requirements	• Script takes input of one already-OCRed pdf file (a dictionary) • Uses regex or other fuzzy search approaches to parse the structure of the file • Returns the words from the dictionary in a structured format (xml, csv or sql) • Structured into "headword", "part of speech, "related text"
Non-essential Requirements	• Script works on files that have not already been OCRed, i.e. runs the files through OCR as part of the pipeline • Script can take multiple pdf files as input • User can pre-select certain filters so that only a subset of the total dictionary information is returned • A simple GUI so it can be used by all users regardless of computer literacy

5. Alternative Solution 1 – Independent Web Application

Solution Description

This solution revolves around the development of a web application, specifically one that is not dependent on external services or connections such as living databases, an important distinction to recognise as it does allow for operation without network connection. This will provide the user with an easy, convenient method to carry out information extraction on dictionaries.

In terms of resources required it is a minimal list as the main item being the project's server host will be covered by Western Sydney University's SCEM department for no operation fee, this is a readily available solution with the added benefit of tailored support and a mutual understanding between the involved parties. Moreover, technologies used are all open source which eliminates any potential licensing costs from operation, specifically these packages are Mozilla's PDF.js for pdf to text conversation, Tesseract.js for image to text recognition, and Datatables by SpryMedia for data presentation and exporting.

Technologies and Programming Language	
Website	HTML5 CSS
Server Side	JavaScript
Data Containers	JSON, XML, CSV

JavaScript Libraries	**Purpose**
PDF.JS	Extract information from an already OCRed PDF file such as: • Word/Text • Height and Width of word/text • X and Y coordinates of each word/text • Font style of the text/word
Tesseract.js	Extract text from an image file
Datatables	Aid with the presentation and outputting of results

Business Case

Benefits of following a web-based solution are far and wide, one being the convivence and accessibility. Given that all the required resources are collated on an external server this cuts the user responsibilities down to the bare minimum, all that is required to operate is a browser and a URL and eliminates any installation process where confusion can sprout as well as availability being the best ability not requiring a specific operating system or prestored local files. In addition, accessibility can also be felt during the user experience as a web program is generally less intimidating to learn as users will think of it as just another website rather than new software to learn.

Furthermore, a web approach offers a large variety of packages and frameworks to integrate during development in turn leading to a more robust and streamlined final project. The advantages in this scenario will be felt by the development team allowing for a quicker timeline leaving more room for added functionality and error correction to provide a better application. Also, should future development be desired specifically pivoting to the inclusion of a living, collective database this solution will more easily accommodate the transition as it is already connected to the web and resources set up.

Moreover, the technologies used in this scenario, these being a hosting server via Western Sydney University, tesseract.js, pdf.js, and Datatables all offer open source use of their technologies meaning there will be no running cost to maintain functionality.

Risks and Constraints

The primary constraint associated with a web application such as this is the inconvenience when it comes to running it in an offline environment. Due to the fact that the product being hosted on an external server it typically requires an internet connection to use, however while it is possible to run offline the process of downloading the developer code and package repositories is quite tedious and alienating for even advanced users.

Furthermore, on the topic of running the program via an external server this creates a reliance on a third party which adds another branch of communication to manage and trust that it remains operational until further notice.

6. Alternative Solution 2 – Desktop Application

Introduction

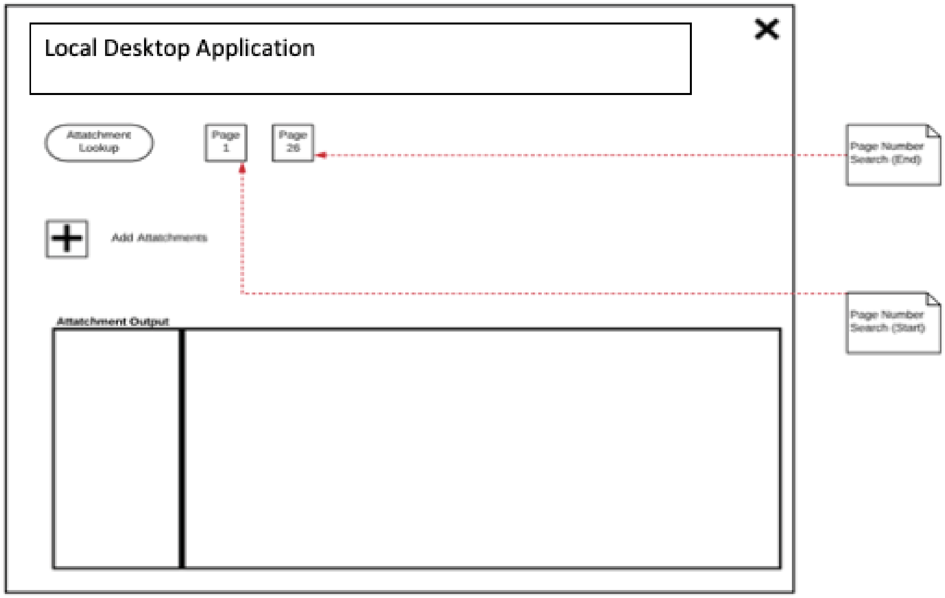
Exceptional software development teams must develop high quality and user-friendly interfaces. Thus, the importance of human computer interaction is vital to a functioning system. In the development phase of our project we must consider all potential users who will interact with this system. This includes properties such as cognitive, behavioral, perception, efficiency and physical aspects must be considered (Cornel University, 2019). Therefore, for reasons given in the following report our project team (KS1912) working on extracting structured information from dictionary PDFs for our client Rachel Hendery and her team of linguists at Western Sydney University have decided to meet project requirements through a local desktop application.

Accessibility

Desktop applications generally use a standard design language and are integrated with the operating system. Using a graphical interface in the program means users can easily identify the common elements such as select file and textbox entries. Explicitly, one of Rachel's non-functional requirements was a GUI interface so other members of her team who are not tech savvy would be able to easily use it. With the development of a visual system, Rachel and her team will be able to access the proposed application instantly via and executable application that imitates a familiar environment in comparison to an exclusive command prompt run script. Users can easily access the application from any Windows based operating system, the most common OS and does not require a niche platform such as Linux or one with a high barrier of entry such as macOS. Moreover, a windows approach can be developed via any system and later adapted to a Unix solution whereas a macOS must be developed on apple devices adding another layer for development teams to accommodate.

Interface

The interface design will be guided closely by Schneiderman's eight golden rules of interface design. Therefore, we will implement simple error handling, consistency in design and reduction of workload through an easy and straight forward design. More details will be discussed in 1.3 System Analysis and Design.



The prototype above illustrates a minimalistic design. We have chosen a simple design as it promotes a system that is straight forward and easy to navigate. We will use instructions for the user using well labelled buttons. Error handling will also be used to further instruct the user on proper use of the application.

Development

This desktop application will not require a database or a dedicated server, it is completely featured in the desktop environment. Development techniques with the aid of C++ for Windows is easy to implement with a local desktop application. A local desktop application will be the simplest technology to use for both client and development team. Moreover, changes and updates are easily implemented in comparison to other technologies.

Risks and Constraints

1. Inconvenient access since it is only accessible to computers that runs Windows as its operating system and does require file installation
2. The user might encounter difficulties in installing the desktop application as it is more complicated to install and run compared to other type of applications.
3. Installing desktop application will require an empty local memory space.
4. Desktop application development takes more time as it is classified as a specialized type of development.

Conclusion

Project requirements will be implemented efficiently with the desktop application. Rachel's requirements will be sufficiently met for not only her, but her team as they will all have access to the application for their own project. Moreover, requirements set can be met through desktop application languages like C++ to output the correct dictionary data. Changes to code and the desktop application can be easily managed, tested and implemented as we progress through developmental stages in our project.

7. Alternative Solution 3 – Mobile Application

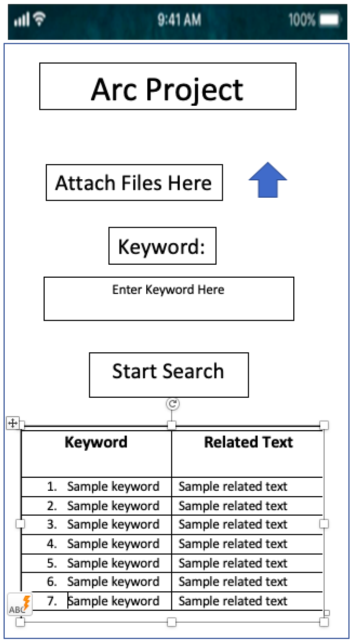
Solution Description

A mobile-based application built with Swift and JavaScript for IOS and Android devices respectively on the frontend and PHP using the Laravel framework on the server side is an appropriate solution for the project. It would allow the user to upload a dictionary PDF file and a keyword search can be performed with the results being presented and the query saved in a database. Dictionary PDF files will also be saved in the database for easy access.

It can be implemented without having to register a domain for it, but a fully qualified domain name can be registered. Essentially, it will only need a hosted space for the uploaded dictionary files and a database within which search results can be saved. In addition, it needs to be registered on Google Playstore and Apple's AppStore to make it accessible for both Android and IOS users.

Technologies and Programming Language	
Android	JavaScript
IOS	Swift
Server Side	PHP
Data Containers	JSON, XML, CSV
Database	MySQL

Sample Mobile Interface:



Keyword	Related Text
1. Sample keyword	Sample related text
2. Sample keyword	Sample related text
3. Sample keyword	Sample related text
4. Sample keyword	Sample related text
5. Sample keyword	Sample related text
6. Sample keyword	Sample related text
7. Sample keyword	Sample related text

Business Case

This solution is an efficient way to automate the searching process. It is compatible with both Android and IOS devices and can be accessed from anywhere. Web hosting is readily available from a number of providers. Moreover, users do not have to purchase a device since most if not all have access or own a smartphone. In addition, accessibility will not be a problem as smartphones are handheld and can be used without any restrictions.

Following on from the topic of accessibility and mobile's ability to be brought everywhere, this opens up the possibility for additional, mobile exclusive features to be added such as taking a picture of a wild dictionary from a colleague or resource collection and have the image be processed by the OCR software and stored in a database for later use.

Risks and Constraints

- There is a cost associated when hosting an application on each device store of $99 USD yearly and $25 USD one time on Apple and Google play stores respectively
- The development time for PHP takes longer as it is more complicated than other server languages.
- Users not using cloud-based storage must deal with inconvenience of cables for file transfer unless accompany web portal is developed.
- As the solution relies on a third-party server, this adds another group to manage and place trust in, any unforeseeable downtime could affect the project development and business use once completed
- Due to the core integration to an online, integrated database the application will require an internet connection for full functionality

8. Recommended Solution with justifications

Alternate solution 1, a localised web application, is the advised course of action going into the future. This solution involves the creation of a website that users can upload dictionaries to and gain an instant output straight from their browser with no additional set up required. To aid the program a series of open source packages are used. Reasons for recommending solution 1 include convivence, development advantages, and low barrier to entry.

The local desktop application, while completely self-sufficient and more powerful in terms of potential, is not deemed suitable for a project of this scope and determined as overboard for the current timeframe and expected functionality. In comparison, alternate solution 1 offers a convenient and accessible process to carry out the data extraction, this is due to the web hosting only requiring a browser to operate at not reliant on a specific operating system architecture or the installation of files and software. By extension, the desktop solution by nature is platform dependent during development and would require editing to run on each of the 3 main systems or at the very least when developed in a universal language such as Python it requires the installation of a compiler which is quite an alienating experience for the common user. Furthermore, in relation to development a website environment offers a wide variety of generic packages with easy access and integrations, whereas on the other hand desktop application due to the added power and more specialised development requires more focus and will take more time to integrate especially for the multiple systems.

When compared to the alternate solution 3 which involves the creation of a mobile application with server integration the advantages of the added technology and potential are unable to outweigh the disadvantages. Similar to the desktop application, when developing for mobile the programs are platform dependent and so to enable a accommodating solution effectively two solutions must be developed due to the wildly different approaches to the program's code in android and iOS. Also, the server and living database are not a required feature and to add feature which it not wanted effectively makes it useless. It is not worth the investment of time spent researching, initialising, implementing, and testing, to work on a feature that may only be used in select cases where as it could be spent fine tuning error correction and aiding the user experience. In addition, with the integration to a server it adds the strict requirement for internet connection whereas in solution 1, albeit tedious to download a website's source files, is still able to work with full functionality in an offline environment. Furthermore, a web application can still be run on mobile so the convivence one might gain from having the software at their fingertips is already available except solution 1 also can be used on their workstation for a more efficient workflow. Not only that, given this project is operating under a budget of zero the cost to publish an application is not planned for.

To summarise, alternate solution 1 a standalone web application, is the recommended course of action for the project at hand. While all solutions will serve the same functions and contain the same features this has the added benefit of accessibility given that it is a remotely hosted website with no installation required, offers a more streamlined development process leading to extra time to fine tune the data extraction and enhance the quality of life, and offers the possibility of further scaling to implement a living database where past results are stored.

9. Ranked High-level Business functions and use cases

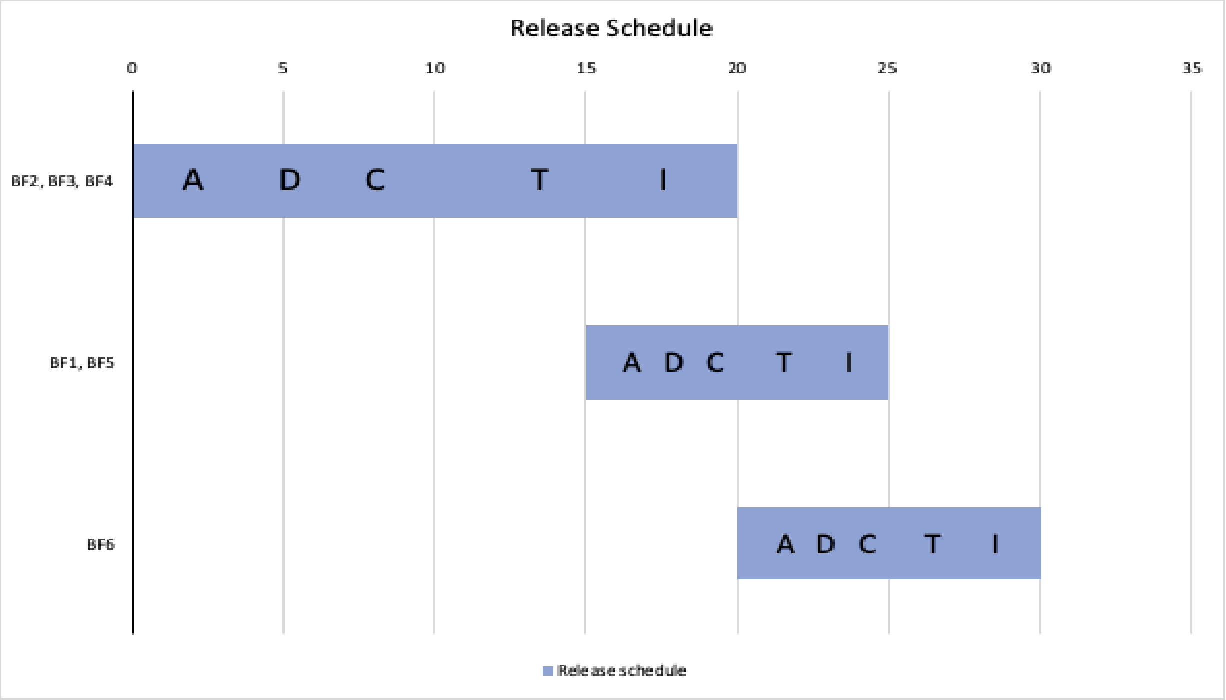
9.1 High-level business functions

BF ID	High Level Business Function	Ranking
BF1	Upload PDF files	Essential
BF2	Search Parameters	Essential
BF3	Error Correction	Essential
BF4	Output file	Essential
BF5	Export file	Essential
BF6	Filter Search Results	Optional

9.2 Use Case List

Function	Type
1. Upload PDF files a. Dynamically select file i. Add a file ii. Change file selection b. Multiple file upload i. View upload list ii. Add a file iii. Remove a file	Essential
2. Search Parameters a. Page search range i. Enter search beginning ii. Enter search ending b. Dictionary features i. Include Example sentence ii. Specify entry breakdown character	Essential
3. Error Correction a. Whitespace size i. Adjust size of whitespace b. Dictionary Columns i. Enter column amount	Essential
4. Output file a. File structure i. Select columns to keep ii. Enter file name	Essential
5. Export file a. Save to Desktop i. Select file format ii. Export file	Essential
6. Filter search results a. Character String i. Enter search string b. Part of Speech i. Add a speech term ii. Remove a speech term	Optional

10. Development Release Schedule



A - Analysis; D - Design; C – Construction; T – Testing; I - Implementation

11. Conclusion

After concluding this document there are several key takeaways including, a breakdown of the project at hand and the groups involved, an insight into the possible development solutions with their associated benefits and risks, as well as a deconstructed view of the program's functionality and related user interaction.

Following on from here once approval is confirmed and all parties are on the same page development will take place at the forefront of the project and before long a solution will be delivered to best suit your needs and aid with further study in linguistic research.