KS1912- Project Abstract Arc Extraction

Problem

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.

Solution Description in Brief

To resolve the given issue, it is proposed that a local web bad application will be developed as it provided excellent accessibility for any member that uses the service as well as a large selection of packages to aid system performance and user quality of life. The system will take a file and relevant search parameters to extract string from an OCR'ed pdf file and display each record in an interactive table to which it can then be exported.

Technologies used

The primary technologies used include PDF.js to help extract the text from a file to then process, utilised the object return, and coordinates keywords and related text can be split up. Following, once split and cleaned the data will be passed to a html table with Datatables applied which allows for user interaction and data manipulation with features such as edit, filter, and export. Finally, Electron has been implemented to transform the application into a localised program one can access anywhere straight from the desktop with the added feature of local storage and independence from the internet.

Client Details

Doctor Rachel Hendery is an Associate professor of Digital Humanities at Western Sydney University, completed her doctorate on the observation of changes in relative clauses constructions cross-linguistically, tying into the historical typology. On the average day Rachel is working as a Linguist looking for ways new technological developments can help in the further study and research of language contact and change in the Pacific. Furthermore, Dr Hendery is involved in the supervision of postgraduate projects regarding digital humanities such as data visualization, mapping, language, virtual reality, simulation as well as more worldly topics including typology, historical linguistics and contact linguistics. Currently this has brought her sights on a 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.

Project Team

This project has been by group KS1912 developed by students Josephine Paculio, Joshua Dib, Roger Bernado, and Caleb Smith, otherwise known as Joroca, during their studies in Professional Experience project. With each member coming from different backgrounds and streams of education surrounding the information technology field it provides the team with a broad vision and a complementary set of skills to serve a complete, robust solution.