

Project Proposal Jacob Evans

by Jacob Evans

Submission date: 20-Oct-2025 02:01PM (UTC+0100)

Submission ID: 265737006

File name: 50439_Jacob_Evans_Project_Proposal_Jacob_Evans_695044_1915280344.docx (135.4K)

Word count: 1805

Character count: 10877

Creating a conversational AI tutor that
allows the user to learn and converse
in the Welsh language

Research Proposal

Submitted by Jacob Evans

Proposed Supervisor: Dr Seena Joseph, James Williams

1500 Words

Contents

1. Introduction 3

2. Background to the Proposed Research..... 3

3. Research Problem statement 4

4. Research Aim & Objectives..... 4

5. Methodology 5

6 Project Plan 6

7. References 7

1. Introduction

The Welsh language is a major importance to Welsh history, nationality and culture. With Wales having a population of 3.1 million in 2024 [1], there is only estimated 538,000 (17%) people who can speak the language in 2023 [2], and this seemingly diminishing every year, there needs to be more programs and applications made in Welsh, as well more applications to learn Welsh so the language can remain prominent in the future.

In recent years, priority for Welsh speaking has diminished in Wales, with Welsh language projects seeing cancellations and funding withdrawn [3]. Evidence of this is projects like “Cymraeg I Blant” in 2016 that was scrapped after 4 years [4]. The aim for this project was to encourage parents to teach and speak Welsh to children, which allowed Welsh families to get together once a month and strengthen the Welsh culture with the children, as well as other Welsh speaking families.

Duolingo being the most popular language learning application, has since stopped development of the Welsh language on the app since 2023, to concentrate on more popular languages. This after the app stating that the Welsh language was one of the fastest growing in the UK in 2020 [5]. When loading the application, it states that there is about 100,000 people learning the Welsh.

The idea of this project would be to promote and provide users with a Welsh language AI that allows them to start study and converse the in the Welsh language. Although 27% of Wales are Welsh speakers, only a small percentage speak Welsh daily.

2. Background to the Proposed Research

With the introduction of the project in mind, we can see that there is a call for this type of research and technology. With the decline in the Welsh medium, and projects being “scrapped”, and the development of AI being on the rise since ChatGPT was released in 2022 [6], there has been a trend in the development of AI, with X releasing their own AI just this year. The advancements of these large language models have transformed natural language processing [7] have allowed computers to create human-like text. With English being the most pre-dominant language in LLM, there is a call for different language models, such as Welsh. This allows for higher accessibility. A telecom company known as Veon states that despite working out of countries such as Pakistan and Ukraine there is under-developed AI tools and LLM in these languages [8]. A research project into developing AI in languages such as Welsh allows us to bridge the “gap in AI language”.

There is an ethical issue regarding this topic, which would be reducing the need for teachers in the future, if AI is able to teach correctly with minimal mistakes.. Neil Selwyn’s book “Should Robots Replace Teachers?” [9] does ask this question, and although AI cannot replicate social, political and moral judgments like a human teacher could, an AI may be better for teaching due to repetition. In the article “Repetition, Difficulty, and AI: Enhancing Vocabulary Acquisition for Second Language Learners”, the major conclusion is that repetition has a major role when learning a second language, although other factors do affect the persons’ learning [10]. The way an AI is taught, is very similar to how we as humans are taught, through repetition and a large range of data [11].

A Welsh AI assistant was published in 2020 from Bangor University [12], where they have researched into creating a LLM for the Welsh language, that allows the user to engage with the AI like Siri. The

research made in that paper has allowed for a successful Welsh AI to be created.. This is incredibly useful for those who already speak Welsh but has very little features that allow a non-Welsh speaker to learn the language, as studying the language is more than using a translator. Macsen uses data from Mozilla [13] which is a crowdsourced data retrieval website that allows users record and submit their voices – this data source will be core for creating Welsh recognition for the Welsh AI tutor application project.

CorCenCC [14] national corpus will also be a core source data for the Welsh language. This is a Welsh library of Welsh words (approximately 11 million words), which will be vital for creating a LLM AI in the Welsh language, as they have sourced from all forms of Welsh literature. This is however considered small compared to other language models, therefore quizzes, questionnaires and interviews may be conducted to retrieve as much data as possible, which should reduce mistakes made by the AI.

By designing and evaluating such a system, this project will explore how conversational AI can help preserve and promote the Welsh language, while also offering insights that could apply to other low-resource languages worldwide, as there is a large call for tutoring apps such as KhanmigoAI, and the use of other AI (CoPilot) within education.

3. Research Problem statement

The problem this project aims to address is the lack of language AI created in the Welsh medium, as well as create a functioning Welsh learning app that considers the different pronunciation and dialect of parts of Wales that will allow high accuracy when conversing with the application, as apps such a “Duolingo” and “SaySomethingInWelsh” do not have these options within their application. This should revitalise the Welsh language.

4. Research Aim & Objectives

1. To analyse and evaluate online literature of the Welsh language and apply this to create an intelligent tutoring system for the AI aspect of the project. We would need a low-resource natural language processing which would allow the AI to create sentences from English to Welsh, and vice-versa. The AI will also learn from voice recordings of different Welsh speakers from across Wales to allow it to understand other Welsh dialects.
2. Develop a mobile application through JavaScript and create an API that would allow for the AI to access other information that we may want to integrate later in the process.
3. The application would have a UI that would allow for text input/output, voice input/output and determine what level of proficiency the user would have (beginner, intermediate, expert) and provide lessons depending on that proficiency level.
4. The application should have log in feature, that allows for the user to create a personal and secure account, wherein the users’ history and previous lessons will be available to return to at a different time, or on a different device.

5. Methodology

This project would be considered a mixed pragmatic research project to develop a Welsh Tutoring AI application. The project would be considered a “Software Development” research strategy, as researching, implementing, designing and assessing the Welsh AI tutor app would follow this case study. This methodology effectively combines technical innovation and pedagogical evaluation, ensuring that the research contributes both to AI for language learning and Welsh language revitalisation.

Data Sources –

- Welsh language data and word library will be required to train the AI for voice recognition, translation and to create lessons in the Welsh medium. Sources such as CorCenCC () and Mozilla Common Voice. These data sources should allow to fine tune the data model; however own collection of voice recordings may also be used to train speech recognition and pronunciation.
- User interaction like quizzes, interviews and questionnaires will also be used to collect data to gain more data on pronunciation and language – North Wales have a similar yet different language to South Wales.

Ethical Considerations-

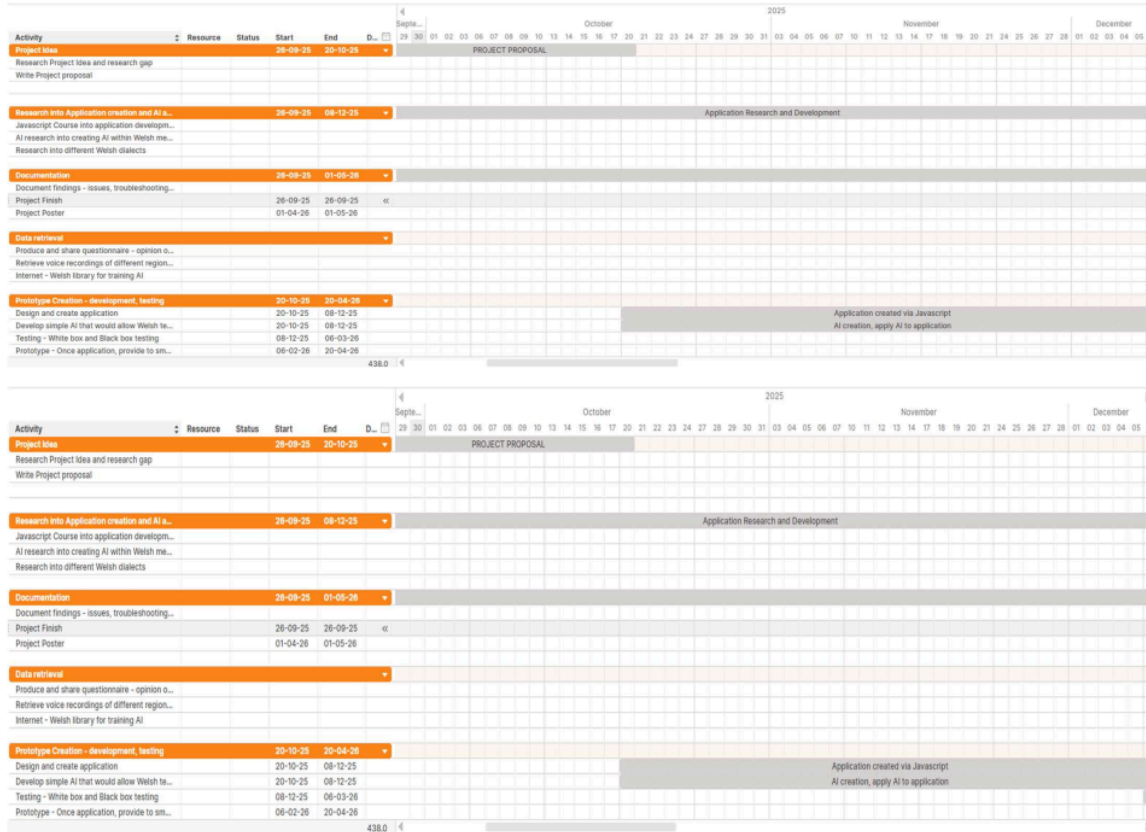
- All participants will need to sign a waiver before the interview is conducted to confirm that their voice will be used to train AI. Participants will need to be advised not to relay any personal information like names, as this is not required and achieve high GDPR compliance. The waiver will also state that the data collected will be deleted within 12 months and the participants reserve the right to have their data deleted at any time.
- The first page of an online quiz/questionnaire will state that if the participant continues, their data can/will be used for the project. By continuing, they are completing a waiver.
- Participation is low risk if these considerations are followed.

Software and Hardware-

- Python and JavaScript are to be used to create the AI and application respectively, with Visual Studio Code being used to develop the application.
- React will be used to build an interactive UI for the application.
- Speechbrain and other AI tutoring libraries will be used to tutor the Welsh AI.
- Audio editors such as Audacity may be used to edit voice recordings/interviews.
- Microsoft Teams to be used for online interviews/questions.
- Home PC with high processing and storage will be used to develop the application and AI.
- Recording devices to be used.

6 Project Plan

Below is a Gantt chart created on Tom's Planner that showcases milestones and estimated dates of when each milestone should be achieved.



7. References

- [1] "Mid year estimates of the population: 2023 [HTML] | GOV.WALES." <https://www.gov.wales/mid-year-estimates-population-2023-html> (accessed Oct 10).
- [2] "Welsh language data from the Annual Population Survey: July 2023 to June 2024 [HTML] | GOV.WALES." <https://www.gov.wales/welsh-language-data-annual-population-survey-july-2023-june-2024-html> (accessed Oct 10).
- [3] M. Mansfield, "Reform's plan to scrap language target 'could threaten Welsh medium education'", Nation.Cymru, <https://nation.cymru/news/reforms-plan-to-scrap-language-target-could-threaten-welsh-medium-education> (accessed Oct 10, 2025).
- [4] A. Shaw. "Welsh language mum and baby project scrapped." <http://www.dailypost.co.uk/news/north-wales-news/welsh-language-mum-baby-project-11061966> (accessed Oct 14).
- [5] J. McCarthy & O. Evans. "BBC- Welsh course to be 'paused' on language app Duolingo" <https://www.bbc.co.uk/news/uk-wales-67197462>
- [6] "ChatGPT" <https://simple.wikipedia.org/wiki/ChatGPT> and <https://chatgpt.com/>
- [7] A. Zubiaga, "Natural language processing in the era of large language models," Front. Artif. Intell., vol. 6, January 12 2024. doi: 10.3389/frai.2023.1350306. Accessed: Oct 10, 2025, [Online]. Available: <https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1350306/full>
- [8] Olivier Sorgho and Alexander Smith, "Veon teams up with partners to bridge online 'AI language gap'," -05-15 2024. Accessed: Oct 11, 2025. [Online]. Available: <https://www.reuters.com/technology/veon-teams-up-with-partners-bridge-online-ai-language-gap-2024-05-15/>
- [9] "Should Robots Replace Teachers?", by Neil Selwyn," Times Higher Education, no. 2436, 2019. <https://www.proquest.com/docview/2447604341?accountid=130472&pg-origsite=primo&searchKeywords=should%20robots%20replace%20teachers&sourcetype=Trade%20Journals>
- [10] L. Zhan, et al., "Effects of Repetition Learning on Associative Recognition Over Time: Role of the Hippocampus and Prefrontal Cortex," Front Hum Neurosci, vol. 12, pp.277, -7-11 2018. doi: 10.3389/fnhum.2018.00277 <https://pmc.ncbi.nlm.nih.gov/articles/PMC6050388/>
- [11] N. A. Prof. Abdullah Al Fraidan, "Repetition, Difficulty, and AI: Enhancing Vocabulary Acquisition for Second Language Learners," Accessed: Oct 10, 2025, [Online]. Available: https://d1wqtxts1xzle7.cloudfront.net/120776013/Repetition_Difficulty_and_AI_1_-_libre.pdf?1736684491=&response-content-disposition=inline%3B+filename%3DRepetition_Difficulty_and_AI_Enhancing_V.pdf&Expires=1760964773&Signature=PZcRwL1VkjyxqQsS6~SSk8JxycX5bYGYov~HcFRrvEucnM52KxAC4mvqmtlmitQzfrAZZc1o0qwFn5R9EU1QfB1O6~CGxp7CjRmzXcbQKN1t7aCKPwUQXuGQHfwdZL4QrWwafxFmMW8KSqAOu~7kq1us3zkVTj2JB7jwZVAeInfZRmSa-4oo6A5vzBvtHqzg31dQ1fbayixR8BTSz1lvXuAi0c9c1K5-

[jZ0htzs-jRbPfxDH6prnAJSh~YsxxwXLltxF~LZZ52~-scyRbUJk04shyGi2vhSxQiqP3-tARyC3NDRBl3tay7TYSSj6k8KzEpYrbkuRBmn46Q0HH7ZA &Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA](https://api.openai.com/v1/engines/gpt-4o-mini/messages?api_key=jZ0htzs-jRbPfxDH6prnAJSh~YsxxwXLltxF~LZZ52~-scyRbUJk04shyGi2vhSxQiqP3-tARyC3NDRBl3tay7TYSSj6k8KzEpYrbkuRBmn46Q0HH7ZA&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA)

[12] D. B. Jones, "Macsen: A Voice Assistant for Speakers of a Lesser Resourced Language," Accessed: Oct 12, 2025, [Online]. Available: <https://aclanthology.org/2020.sltu-1.27.pdf>

[13] CommonVoice Mozilla - <https://commonvoice.mozilla.org/en/datasets>

[14] "CorCenCC – National Corpus of Contemporary Welsh", CorCennc.org, <https://corcenc.org/> (accessed Oct 12, 2025).

Project Proposal Jacob Evans

GRADEMARK REPORT

FINAL GRADE

GENERAL COMMENTS

72/100

PROJECT PROPOSAL – ACCA6020 INDEPENDENT PROJECT

Student Evans
Jacob **Mark** 72

Intro / A&O (400 words)		
Project title; Identification of a specific problem to be investigated; Well-defined aims and SMART objectives; Explaining why this project is worth doing; Discussion of what the project needs to do in order to be a success; Description any deliverables that will be produced at the end of the project	20	16
Clear project title.		
Identification of a specific problem to be investigated and worthiness & value of the project is clearly outlined.		
The aims and objectives are well-defined.		

- Existing objectives need to be refined. 1/2 additional objectives covering user testing and evaluation could be documented. Discussion of what the project needs to do to be a success is analysed and documented. Intended deliverables that will be produced is documented.		
Background to the Project / Literature Review (400 words)		
Description of how the work to be done by in the project relates to the work identified in the literature review; Discussion of any ethical issues raised by the project Identifying what has already been done within the field; Topic familiarity; Brief literature review using appropriate sources – 6-8 journal papers and book chapters rather than just websites; Selection of relevant models & concepts Description of how the work to be done relates to some identified research work is documented and cited. Discussion of ethical issues raised by the project are somewhat identified and considered. – storage and use of personal information (privacy) needs some additional consideration.	20	16

There is some identification of what has already been done within the field, cited and referenced. Brief literature review using some appropriate sources. Use of up-to-date and unbiased citations in 1/2 areas could be extended.		
Project Design (500 words)		
Clear statement of which development methods are to be used and why they are appropriate; Clear statement of which tools are to be used and why they are appropriate; Clear description of how data collected will be analysed A research method is clearly documented. – A software development method could also be documented. Statement of which tools to be used and why they are appropriate is documented. - Description of how data will be collected and analysed and which tools will be used for this purpose could be documented.	30	20

Resources / Planning (200 words)		
List of any deliverables; Clear diagram showing both timing and resource use; Completed ethics form; Limited and relevant appendices Deliverables stated. Diagram showing timing is documented. – Easier to read schedule could be documented in table format. Ethics form completed.	15	10
Document Presentation		
Correct spelling, grammar and punctuation; Flow of narrative/fluency; Coherence; Word processed ; Correct use of referencing with bibliography; Strict word limit Good structure, format and presentation.	15	10

- Minor grammar errors in places. - Gantt chart difficult to read. Overall a very good project proposal submission.		
TOTAL	100	72

PAGE 1
PAGE 2
PAGE 3
PAGE 4
PAGE 5
PAGE 6
PAGE 7
PAGE 8