



COMP4801 Final Year Project

Interim Report

Learning Platform for Minority and/or Endangered Languages

January 2025

Member: Chen Jun (junchen4@connect.hku.hk)

Supervisor: Choi Yi King (ykchoi@cs.hku.hk)

Abstract

Many spoken languages across the world are at risk of decline. At the same time, Hongkongers are generally not knowledgeable about such conditions, and language-learning platforms generally only offer courses on major languages such as Chinese, French or Spanish. The platforms that do offer such courses created by the community are now rolling back their community support, which thus calls for a platform with strong community support, dedicated to learning lesser-known and/or endangered languages.

The project aims to provide a platform for those in Hong Kong who are interested in learning various endangered or minority languages around the world, to raise awareness regarding said languages in wider society, and give a voice to language enthusiasts who may have encountered issues promoting their own languages on other platforms. This will be realized using a web application coded using the MERN Stack. Due to technological limitations, languages with vertical writing or sign languages cannot be covered by the platform. In addition, the limited time also does not permit a mobile version of the platform to be developed and launched simultaneously.

In terms of current progress, the user database and prototype design are being worked on, while the first language to be added to the platform has been decided. Challenges currently faced during development include the incorrect display of characters and glyphs, platform design, as well as data collection & processing. In the near future, the base design of the prototype will be finished, to allow for commencement of user interface (UI) development.

Acknowledgement

The author would like to thank Dr. Choi Yi King for providing early comments on the scope of the project.

Table of Contents

Abstract	i
Acknowledgement	ii
Table of Contents	iii
List of Figures	iv
List of Tables	v
1. Introduction.....	1
1.1. Background	1
1.2. Scope of this report	1
1.3. Objectives & aims	2
1.4. Deliverables	2
1.5. Outline.....	2
2. Literature Review.....	3
3. Methodology	5
3.1. Platform features	5
3.2. Technology stack	6
3.3. Data collection	7
3.4. Limitations	8
4. Findings & Discussion.....	9
4.1. Current objectives & progress	9
4.2. Current difficulties	10
5. Project Schedule.....	12
6. Conclusion	13
7. References.....	14
8. Appendix.....	17
8.1. Table 1	17

List of Figures

Fig. 3. 1. The macro structure of the language-learning platform, using the MERN stack for the core infrastructure and a REST API to retrieve data from Wiktionary.	7
Fig. 3. 2. An entry in an English-Pannonian Rusyn dictionary [14], with text in both Latin and Cyrillic scripts. This is contrasted with a search result on archive.org for the same entry in the dictionary, showing the Latin-only encoding used for both Latin- and Cyrillic-script text.	8
Fig. 4. 1. The preliminary database diagram, styled in a more typical SQL style, including variables for different classes.....	10
Fig. 4. 2. The letter “tje” displayed on Wiktionary. On the right, the actual shape of the letter, as used in the Khanty language [21]	11

List of Tables

Table 5. 1. A table describing the schedule of the project, where relevant to current discussion. 12

Table 8. 1. A table describing the schedule of the project. 17

1. Introduction

1.1. Background

Around the world, more than 7000 languages are spoken [1] by numerous groups. However, many of those languages are slowly being spoken less. Ethnologue estimates that 44% of languages - more than 3000 in real terms - are endangered today [2] .

In Hong Kong, societal and familial influence is perceived as more important to learning languages rather than self-motivation. In addition, students learning foreign languages are largely motivated to do so because they hope to achieve greater career prospects from it, or to emigrate to a country where that language is spoken. Less thought is given to self-betterment or improving one's knowledge of geopolitics and global affairs through linguistics [3] . In addition, Hongkongers generally do not speak many languages beyond the East Asian sphere [4] , showing their relative lack of linguistic knowledge in other regions of the world.

On the other hand, most language-learning platforms target people who are seeking to learn major languages, as platforms such as Babbel only offer courses in national languages [9]. These platforms therefore do not sufficiently highlight the situation faced by many endangered languages across the world. Furthermore, as other language-learning platforms such as Duolingo and Memrise grow, they have slowly strayed away from the principle of community contribution, thus leaving language enthusiasts with fewer places to contribute their local knowledge.

1.2. Scope of this report

This report will describe the features of the minority-language-learning platform, as well as the technology and procedures required to create the platform. It will also delve into the challenges and limitations of this project, that is, what the platform will not be able to achieve. The current findings and progress will also be discussed in the report, and the overall project schedule will also be presented. To conclude, a recommendation will be made on the next steps to be taken.

1.3. Objectives & aims

The objectives of this project are:

- To use MongoDB to create a database of vocabulary and word attributes
- To create a database for storing users and their respective attributes and classes
- To create a visually appealing user interface using React
- To develop a minority-language-learning web application using the MERN Stack

On a more abstract level, the project seeks to aid those looking to broaden their horizons through linguistics. Furthermore, the platform encourages community contributions through creating lessons for various minority languages, and aspires to eventually establish a community of minority-language learners and enthusiasts in Hong Kong and beyond.

1.4. Deliverables

This project will deliver a minority-language-learning platform in the form of a web application, which will include features such as word-matching and a community platform to submit languages.

1.5. Outline

The remainder of this report proceeds as follows. Firstly, a literature review will be provided in Section 2, further examining the points made in the background section. Next, some features of the platform, the technology used in the project, as well as how data may be collected for use within the platform will be presented in the methodology in Section 3. Anticipated limitations will also be explored. Then, current findings will be presented in Section 4, that is, the challenges faced at the current phase of the project. Afterwards, Section 5 will review the current progress made on the project, including a plan for the remaining work and the tentative schedule to complete said work. To conclude, the next immediate steps will be presented in Section 6.

2. Literature Review

This section will review some existing literature relevant to the project. As mentioned in the background in Section 1.1, more than 40% of languages worldwide are at risk of decline or extinction.

However, this is not something that particularly concerns people in Hong Kong, as a 2005 study on tertiary students in Hong Kong finds that of the students learning French, German and Japanese, two main motivational dimensions for learning those languages were integrative (i.e. aiming to be like a native speaker or integrate into their culture) and instrumental (i.e. gaining success in work). For Japanese learners, the affective dimension was ranked above the integrative and instrumental dimensions, however the latter two also had a higher mean score than those two dimensions for French and German. Among the motivational dimensions included, two other ones were “macro-context”, referring to the language attitudes of the wider community (within Hong Kong) and ethnolinguistic identity, as well as “micro-context”, that is the language attitudes of peers and parents. It was found that with learners of French, German and Japanese, micro-context ranked above macro-context. In all three languages, the lowest-ranked dimension was “linguistic self-confidence” [3]. In addition, another study was conducted from 2014 to 2015 on the use of spoken languages among Hongkongers. Out of over 2000 respondents, while over 60% reported usage of English, Putonghua and Cantonese, less than 7% reported speaking other Chinese dialects, 5.2% spoke Japanese, 1.8% spoke French, 1.2% spoke Indonesian, and less than 1 percent reported speaking any other languages [4]. This shows that people in Hong Kong in general are neither fluent in many languages nor very interested to learn new ones for non-career purposes.

This could theoretically be remedied, as there are numerous language-learning applications on the market today, either in the form of mobile applications, web applications, or both. For instance, Duolingo is a language-learning application that was first publicly released in June 2012. [5] The platform is identified by a unique light green and white colour scheme, with a mascot owl named “Duo”. Since its launch in 2012, it has quickly become by far the most popular language-learning application [6]. Notably, the platform has garnered much attention

due to its features seeking to “gamify” language-learning - with experience points, leagues, and lives which are lost when an incorrect answer is inputted, Duolingo seeks to portray learning languages as an online game. The languages featured in Duolingo are rather extensive, including many major national languages, as well as some lesser-spoken languages like Yiddish and Navajo, and even fictional languages like High Valyrian and Klingon. However, in general, Duolingo covers widely-spoken languages, mostly national languages such as French or Hindi. In 2013, the “Duolingo Incubator” was launched, where communities of speakers of a given language could come together to work on a new language course to then be pushed to the main application. However, in March 2021, this feature was removed, citing alignment with CEFR standards and the issue of making money off of volunteer-produced courses [7].

Another popular language-learning platform is Rosetta Stone. It is comparatively much older than Duolingo, having been established back in 1992 [8] , and is a helpful application for people seeking to learn major languages while getting an insight into the grammar, largely not provided in Duolingo. Languages offered include only national languages, such as Spanish or German. Babbel is a language-learning platform that was launched in 2007 in Germany. Many forms of online language education are provided, from podcasts to online small-group learning sessions [9]. Like Rosetta Stone, Babbel only offers courses in national languages. An internal group of 200 experts are responsible for creating the courses, and no volunteer-led material is used. In addition to Duolingo, one other platform that provides volunteer-made courses is Memrise, which even in 2012 had already accrued materials from over 100 languages [10]. Languages featured on the platform range from the Lingala language in sub-Saharan Africa, to indigenous languages from Taiwan, the USA, and Sweden. Memrise however also announced the closing of their community forums in late 2023 [11], and that community-created courses would move to an entirely separate website [12] .

All of the above reveal three major issues in Hong Kong and language learning: firstly, people in Hong Kong are not interested in learning languages in general; secondly, widely available language-learning platforms largely only offer courses in languages that are already widely-spoken, and thirdly, even platforms which offer courses in minority languages have limited

community involvement in their courses. All of this is occurring while more and more languages worldwide are in decline or even at risk of extinction, thus raising the need for a dedicated language-learning platform for minority languages.

3. Methodology

In the following section, the methodology with which this project will be realized will be explained. Specifically, in section 3.1, the features of the platform will be given. Section 3.2 will discuss the technology used to create the platform, while section 3.3 will provide further details on how the data will be collected for different languages. Section 3.4 lists several limitations with regards to creating this project.

3.1. Platform features

There will be four major features included in the platform, the first of which is word-matching. That is, a word in the minority language will be provided, and the user must match with it the correct English definition in order to proceed with each lesson. This will be the primary means of language education on the platform, although occasionally short sentences may be given. For the first few times a word appears, an accompanying image may be provided to hint at the English definition of the word, but afterwards a user must remember the meaning of said word. Where applicable, the word-matching may also come with relevant cultural facts.

This leads into the discussion of our second feature, which is an experience points (XP) and level-up system. For each lesson that a user does, they will receive a certain number of experience points for completing that lesson. If a culturally significant word appears in said lesson, the user may receive a larger amount of experience points.

Once a user accrues a certain amount of XP, they will be able to level up, which may enable them to a selection of rewards. One such reward could be the third major feature - a personalized dictionary. This dictionary would store the words that have been learned by the user thus far, and

would aid the user in memorizing the meaning of words. For the aforementioned culturally significant words, a description will be provided for said word, in addition to its definition.

The final planned major feature is an interface for speakers of lesser-known languages to submit their own language to the platform. Thorough inspection will take place before a language is released publicly onto the platform to prevent vandalism or otherwise malicious behaviour. In addition, as words are sourced through Wiktionary, the language must have at least 200 entries on Wiktionary to qualify for approval and use on the platform. This minimizes the chance of under-attested languages being added to the platform when the documentation is not yet complete.

3.2. Technology stack

The MERN stack [13] was chosen to implement the platform. That is, MongoDB is used to create and host the database; Express is used for dynamically rendering the HTML page by passing arguments to templates; React is used to design the front-end architecture, as it provides flexibility and numerous modules for designing an aesthetically pleasing website; and Node.js is used to provide an interface between front-end and back-end architecture, and allows for both sides to be coded with JavaScript. The combination of the four technologies in the MERN stack allow for creating visually appealing web applications with interactive features, and is widely used across the web development field for a variety of projects. In addition, a REST API is used to retrieve words and information about said words from Wiktionary, which will be further discussed in the next section. The overall structure of the platform can be seen in Fig. 3. 1.

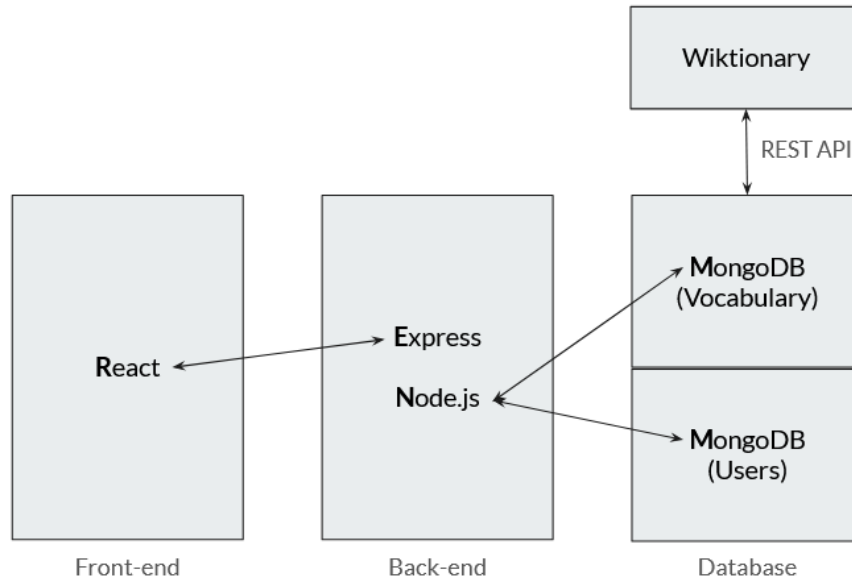


Fig. 3. 1. The macro structure of the language-learning platform, using the MERN stack for the core infrastructure and a REST API to retrieve data from Wiktionary.

3.3. Data collection

For this project, the bulk of the vocabulary and other data comes from Wiktionary. It is one of the largest databases containing words from numerous languages and dialects. While countless resources exist which document various languages, many have not been digitized properly, and are thus challenging to import into a database for the immediate purposes of the project. For instance, Latin encoding may be applied to a dictionary which contains both terms in the Latin script and terms in the Cyrillic script, which makes it impossible to search for terms written in the Cyrillic script, as seen in Fig. 3. 2.

water /'wɔ:tə(r)/ 1. м. вода; **by** ~ по води; **high** ~ плима; **low** ~ осека; 2. *прикм.* водови, водни; 3. *прех.* помачац, намачац, мачац, овлажиц, овлажовац; пирскац; заляц, залівац; наводніц, наводньовац; поляц, полівац; преляц, прелівац; 4. напоїц, поїц; 5. розблажиц, розблажовац; розводніц, розводньовац;

Page 371

water /'wa:ta(r)/ 1..m. Boxa; by "TIO BOAM; high - numa; low - ocekKa; 2. NPUKM. BO/ [OBH, BO1HM; 3. *прех.* TOMAayal, HaMayal, Mayal, OBJIA>HMU, OBJIAMOBAIL; MIMPCKAal; 3aJ4AH, 3AJMBal; HaBOAHil, HaBO/AHbOBal; HMOJIAH, MONIBAL; Tnperan, npeiBan; 4. Hanoi, noir; 5. pO30JlaKk MH, pO36JIAMOBAN; pO3BO/Hil, PO3BOAHBbOBal1;

Fig. 3. 2. An entry in an English-Pannonian Rusyn dictionary [14], with text in both Latin and Cyrillic scripts. This is contrasted with a search result on archive.org for the same entry in the dictionary, showing the Latin-only encoding used for both Latin- and Cyrillic-script text.

As time is limited for the project, Wiktionary is thus the most readily accessible database for a variety of language resources. Words and phrases can be downloaded and imported into the database using the Wiktionary API [15] and Wikimedia REST API [16]. From there, the words will then be categorized into different parts of speech, and stored separately in the MongoDB database.

3.4. Limitations

One major limitation to the project is that with the relatively limited time, it is not possible to add a substantial number of languages to the application. For this project, a policy of “quality over quantity” will thus be implemented; only a handful of languages will be added to the platform, but extra attention will be paid to said languages to ensure that the language profiles are both accurate and interesting to read. The language submission feature will also permit for more languages from more parts of the world to be added to the platform.

The second limitation is that sign languages cannot be covered with the platform. This is also a limitation encountered by the aforementioned major language learning platforms.

The third limitation is that it is not possible within the allotted project development time period to create a mobile equivalent of the platform. If such an equivalent were to be made, it would have to go through a nearly entirely different design process, and both forms would have to undergo user testing before being pushed out at the same time for the project, which is not practical for the time provided at this stage. However, a mobile version presents an opportunity for further development of this project after the final presentation, and will be looked into.

4. Findings & Discussion

In this section, the current findings of the project will be presented and discussed. Section 4.1 discusses the objectives currently being handled, and the progress that has already been made, whereas section 4.2 will lay out the difficulties being encountered at this stage of the project.

4.1. Current objectives & progress

The current objectives being tackled are to create a database for storing users and their respective attributes and classes and to create a visually appealing user interface using React. It is expected that once these two objectives are finished to the current expected extent, a database with detailed parameters for different classes of users will be created and fully connected to the back-end, and, for the current stage, a workable prototype of the user interface will be made and presentable as a bare-bones product prior to finetuning.

In terms of progress being made, the technological component, split into detailed stages, will be discussed further in Section 5.1. In terms of actual content on the platform, at least one language has been confirmed to be added to the platform: Macanese. Macanese is a language formed from the combination of Portuguese, Malay, India and Cantonese influences, and is spoken by a small Macanese minority in the city of Macau today, as well as a Macanese diaspora abroad in Portuguese-speaking countries such as Brazil or Portugal [17]. Occasionally, plays are still performed in the Macanese language, such as “Vêm di volta pa jantâ”, which was performed in November of 2024 [18]. However, it is estimated that fewer than 20 native speakers of Macanese remain today, most of them being elderly people. This language has been chosen to add to the

platform because taking into account that the primary demographic for this platform is Hongkongers, Macau has a similar cultural background to Hong Kong, and Macanese itself borrows quite significantly from Cantonese which is spoken by almost all Hongkongers. Other well-documented minority languages, such as Kildin Sámi or Pannonian Rusyn, may be considered for future additions.

In addition, Fig. 4. 1. provides the preliminary diagram for the database. The “class” attribute in ‘user’ indicates whether a given user is a regular registered user, or an admin. For the ‘word’ database, “lemma” is ‘a word considered as its citation form’ [19], or in other words, the base form of the word. For instance, in English, “is”, “am”, “are” are not lemmas, while “be” is a lemma. Thus, the Boolean “is_lemma” records whether a word is in its base form, and if not, the inflection is recorded in the “inflection” attribute. “lang_id” refers to the language the word is in, using a two- or three-letter code as set out on Wiktionary, for instance “en” for English. “part_of_speech” refers to whether the word is a noun, verb, adjective, or something else.

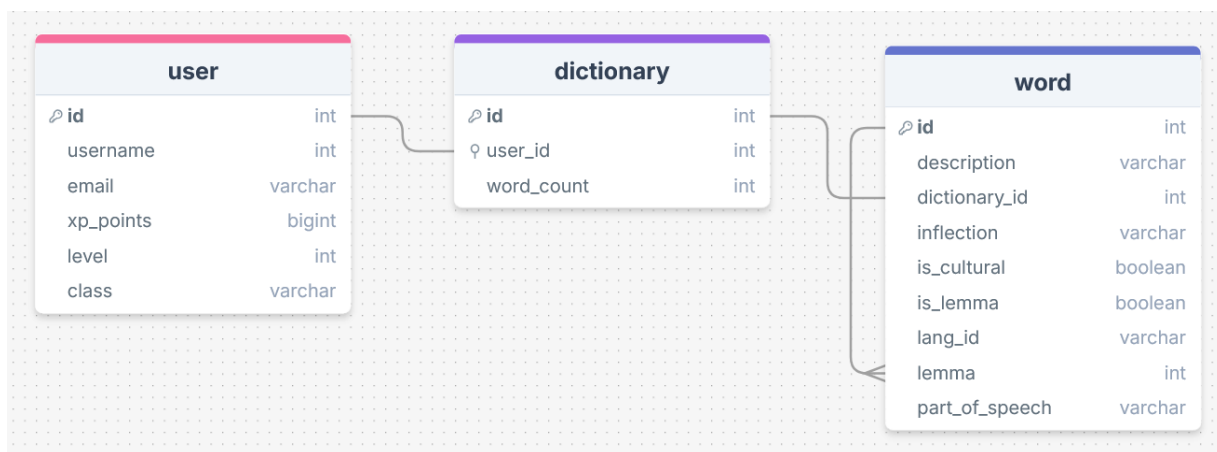


Fig. 4. 1. The preliminary database diagram, styled in a more typical SQL style, including variables for different classes.

4.2. Current difficulties

The main difficulty currently relates to displaying languages. There are numerous scripts used to write languages across the world: Latin, Cyrillic, Greek, Arabic, and so on. Some of these scripts have not been encoded into Unicode yet, and so those languages may only be covered in passing

due to technological limitations. Even if the scripts themselves are displayable on electronic screens, there may be individual characters which may not be supported. For example, the Cyrillic “tje” letter, used in the Khanty language, cannot be displayed on Wiktionary even on modern computers running Windows 10 at the time of writing this report [20], as can be seen in Fig. 4. 2.

Eastern Khanty [\[edit\]](#)

Letter [\[edit\]](#)

Ყ (t') (*upper case* Შ)

1. A *letter of the Eastern Khanty alphabet*,



Fig. 4. 2. The letter “tje” displayed on Wiktionary. On the right, the actual shape of the letter, as used in the Khanty language [21] .

This issue is only exacerbated with antiquated systems. In addition, some typefaces, created with English in mind, may not support accented letters, so caution will be taken when choosing a typeface to be used across the platform. Otherwise, display issues may occur if the wrong typeface is selected.

Where scripts are able to be displayed, there may be other encoding issues, such as text direction. For instance, the Arabic and Hebrew scripts go from right to left, while the traditional Mongolian script is written top-down. To facilitate text display, for this project, languages written using top-down scripts or scripts not supported by Unicode will not be featured as learnable languages. Furthermore, fonts from the Noto Sans [22] family will be selected for display on the platform, as they provide extensive support with a large number of glyphs, especially in Latin, Cyrillic and Greek scripts. If a language using another script, such as Arabic, is to be added to the platform, a different font will have to be used for that language.

The second difficulty is platform design. It is important to strike a balance between usability, simplicity, and retention, that is, users will not immediately click off of the page and not return

after the first time seeing it. To create such a design, Nielsen’s heuristics [23] can be employed. Nielsen’s heuristics are a set of guidelines for creating a page with good usability. This takes into account, for instance, the tendency for users to have a Z-shaped reading pattern: going left to right on the first line, diagonally returning to the left before resuming this process on the next line. All in all, by following Nielsen’s usability heuristics, it can create a page which is both aesthetically pleasing and enjoyable to use.

The third difficulty is the collection and processing of data, especially the latter. As mentioned earlier, the Wiktionary REST API will be used one time to collect words from Wiktionary, such that the user does not have to access Wiktionary every time they begin a lesson in a given language. If a large change occurs on Wiktionary with regards to the relevant language, the API will be used again to retrieve the most up-to-date version of the vocabulary. However, the processing of such words is not straightforward, and there does not exist a one-size-fits-all solution to adapt to the needs of this platform. Not all attributes of a word are necessary for use for the purposes of the platform.

5. Project Schedule

In the upcoming section, the current work plan and overall project schedule will be explained and reviewed. A relevant section of the schedule will be displayed here. The overall schedule may be accessed in the Appendix.

Table 5. 1. A table describing the schedule of the project, where relevant to current discussion.

Date	Tasks, Milestones, Deliverables	Progress
1st October 2024	Project Plan	Complete
October 2024 - 12th January 2025	Begin coding, set up foundations for the platform	In progress
15th January 2025	First presentation	Complete

26th January 2025	Preliminary implementation	In progress
	Interim Report	Complete
27th January - 20th April 2025	- Finish coding main sections - Finalize design of platform - User testing - Deployment	Not started

As seen in the schedule, the phase “set up foundations for the platform” as well as the preliminary implementation are currently still in progress. As these phases were expected to be finished by the submission of this report, the project is thus behind schedule.

6. Conclusion

This report presents a learning platform for minority and/or endangered languages, given the current sociopolitical context where a sizeable percentage of languages are at risk of decline or extinction, in addition to the general low level of knowledge or interest regarding global languages across Hong Kong society as a whole. In addition, existing language platforms are overly focussed on teaching national and/or widely-spoken languages, and the platforms that do provide community-run minority language courses have been rolling back their community support as of late. This thus creates an opportunity for a dedicated learning platform for languages that are endangered or going into decline, which among other features, provides for speakers of said languages to submit their own languages and courses onto the platform.

The project aims to provide a platform for those in Hong Kong who are interested in learning various endangered or minority languages around the world, to raise awareness regarding said languages in wider society, and give a voice to language enthusiasts who may have encountered issues promoting their own languages on other platforms. In more concrete terms, the project is being realized through a web application coded using the MERN Stack, including a database for

users and a database for vocabulary created using MongoDB, and an attractive user interface using React. Due to technological limitations, languages with vertical writing or sign languages cannot be covered by the platform. In addition, the limited time also does not permit a mobile version of the platform to be developed and launched simultaneously.

In terms of current progress, the user database and prototype design are being worked on, and it has been decided that Macanese will be the first language added to the platform. It has so far been found that there may be issues with displaying different letters from various languages, as some widely available fonts do not support accented letters. To alleviate this issue, the Noto Sans font family from Google will be used. Another difficulty is to balance usability and aesthetics in platform design, for which Nielsen's usability heuristics can be adhered to. Processing of data collected from Wiktionary is another difficulty, as there does not exist a template for handling the desired data for future use. Looking to the immediate future, it is recommended that both the user and lexicon sections of the MongoDB database as well as the base design of the platform be finished, such that vocabularies and lexicons can commence being imported into the database, and the visual design of the user interface can begin.

7. References

The following is the list of references.

- [1] "How many languages are there in the world?," Ethnologue, <https://www.ethnologue.com/insights/how-many-languages/> (accessed Oct. 14, 2024).
- [2] "How many languages are endangered?," Ethnologue, <https://www.ethnologue.com/insights/how-many-languages-endangered/> (accessed Oct. 14, 2024).
- [3] G. Humphreys and M. Spratt, "Many languages, many motivations: A study of Hong Kong students' motivation to learn different target languages," *System*, vol. 36, no. 2, pp. 322, Jun. 2008. doi:10.1016/j.system.2007.09.010

- [4] J. Bacon-Shone, K. Bolton, and K. Luke, “Table 3.1 Knowledge and use of spoken languages by percentage”, in *Language Use, Proficiency and Attitudes in Hong Kong*, Hong Kong: Social Sciences Research Centre, the University of Hong Kong, 2015, pp. 18
- [5] S. Loeb, “When duolingo was young: The early years,” VatorNews, <https://vator.tv/news/2018-06-22-when-duolingo-was-young-the-early-years> (accessed Oct. 15, 2024).
- [6] Straits Research, “The best language learning apps of 2024 by downloads.,” Straits Research, <https://straitsresearch.com/statistic/the-most-popular-language-learning-apps> (accessed Oct. 15, 2024).
- [7] Matt, “What happened to the Duolingo Incubator?,” duoplanet, <https://duoplanet.com/what-happened-to-the-duolingo-incubator/> (accessed Oct. 15, 2024).
- [8] “About Us,” About Rosetta Stone® | Company History and Mission, <https://www.rosettastone.com/about/company/> (accessed Oct. 15, 2024).
- [9] Babbel GmbH, “It starts here,” Babbel, <https://www.babbel.com/about-us> (accessed Oct. 15, 2024).
- [10] D. Baker, “Crowdsourced Mnemonics: How Ed Cooke teaches languages with memrise,” Wired, <https://www.wired.com/story/crowdsourced-mnemonics/> (accessed Oct. 15, 2024).
- [11] “Closure of the MEMRISE Community Forum,” Memrise Classic, <https://memrise.zendesk.com/hc/en-us/articles/20736101603473-Closure-of-the-Memrise-Community-Forum> (accessed Oct. 14, 2024).
- [12] Memrise Team, “Major update: A new version of the app is coming!,” Memrise, <https://www.memrise.com/blog/major-update-a-new-version-of-the-app-is-coming> (accessed Oct. 15, 2024).
- [13] “Mern stack explained,” MongoDB, <https://www.mongodb.com/resources/languages/mern-stack> (accessed Oct. 15, 2024).

- [14] Fejsa, M., Šlemender, M., Čel'ovski, S. “water”, *Английско-руски словник* [English-Rusyn Dictionary], Novi Sad: Faculty of Philosophy, 2022, [pp. 371](#)
- [15] “MediaWiki API help - wiktionary, The free dictionary,” Wiktionary, <https://en.wiktionary.org/w/api.php> (accessed Oct. 15, 2024).
- [16] “REST API documentation,” Wiktionary, The free dictionary, https://en.wiktionary.org/api/rest_v1/ (accessed Oct. 15, 2024).
- [17] S. Reporter, “Macau’s ‘sweet language’ on verge of disappearing,” The Mail & Guardian, <https://mg.co.za/article/2011-11-02-macaos-sweet-language-on-verge-of-disappearing/> (accessed Jan. 20, 2025).
- [18] TFAdmin, “戲劇農莊【黑盒劇場演出計劃2024】 第三回 《回家吃飯 Vêm Di Volta Pa jantâ》,” 戲劇農莊, <https://www.tfarmers.org/2024/11/19/%e6%88%b2%e5%8a%87%e8%be%b2%e8%8e%8a%e3%80%90%e9%bb%91%e7%9b%92%e5%8a%87%e5%a0%b4%e6%bc%94%e5%87%ba%e8%a8%88%e5%8a%832024%e3%80%91-%e7%ac%ac%e4%b8%89%e5%9b%9e-%e3%80%8a%e5%9b%9e%e5%ae%b6%e5%90%83/> (accessed Jan. 20, 2025).
- [19] “Lemma definition and meaning,” Collins English dictionary, <https://www.collinsdictionary.com/dictionary/english/lemma> (accessed Jan. 20, 2025).
- [20] “□ - wiktionary, The free dictionary,” Wiktionary, <https://en.wiktionary.org/wiki/%E1%B2%8A> (accessed Oct. 15, 2024).
- [21] Kwamikagami, “File:unicode 0x1C8A.SVG,” Wikimedia Commons, <https://commons.wikimedia.org/w/index.php?curid=152543159> (accessed Oct. 15, 2024).
- [22] “Noto sans,” Google Fonts, <https://fonts.google.com/noto/specimen/Noto+Sans> (accessed Nov. 30, 2024).
- [23] “Usability principles – Jakob Nielsen’s 10 usability heuristics for user interface design,” UX24/7, <https://ux247.com/usability-principles/> (accessed Nov. 30, 2024).

8. Appendix

8.1. Table 1

Table 8. 1. A table describing the schedule of the project.

Date	Tasks, Milestones, Deliverables	Progress
September 2024	• Set up project website • Plan project • Make first contact with supervisor	Complete
1st October, 2024	Project Plan	Complete
October 2024 - 12th January, 2025	Begin coding, set up foundations for platform	In progress
13th - 17th January 2025	First presentation	Complete
26th January 2025	Preliminary implementation	In progress
	Interim Report	Complete
27th January - 20th April 2025	• Finish coding main sections • Finalize design of platform • User testing • Deployment	Not started
21st - 26th April 2025	• Finalized Tested Implementation • Final Report	Not started

	• Final Presentation	
26th - 29th April, 2025	• Record exhibition video • Update project website with deliverables	Not started
30th April 2025	• Project Exhibition • 3-minute video	Not started