Documenting Endangered Languages with *LangDoc*: A Wordlist-Based System and A Case Study on Moklen

Piyapath T Spencer
Faculty of Arts, Chulalongkorn University
linguistics@piyapath.uk

Abstract

Language documentation, especially languages lacking standardised writing systems, is a laborious and time-consuming process. This paper introduces *LangDoc*, a comprehensive system designed to address challenges and improve the efficiency and accuracy of language documentation projects. LangDoc offers several features, including tools for managing, recording, and reviewing the collected data. It operates both online and offline, crucial for fieldwork in remote locations. The paper also presents a comparative analysis demonstrating LangDoc's efficiency compared to other methods. A case study of the Moklen language documentation project demonstrates how the features address the specific challenges of working with endangered languages and remote communities. Future development areas include integrating with NLP tools for advanced linguistic analysis and emphasising its potential to support the preservation of language diversity.

1 Introduction

Amongst the very first tasks in language documentation are to collect and record vocabulary of the language. Traditionally, language data have been collected and stored in its most primitive form, often involving manual recording on paper or default word lists, sometimes with audio recording. This process is yet the most gruelling and labour-intensive. Despite the use of technology and/or computer-assisted systems in latter studies (e.g. Black and Simons (2006), Yooyen (2013), Dunham (2014), van Esch et al. (2019)), the heavy reliance on humans is inevitable, especially converting field notes into computer-stored data prior to any further analyses.

Human errors normally weaken the efficiency of a documentation project and contributes to various issues within the system (cf. Rasmussen and Vicente (1989), Compton (2014)), including compromising the overall quality of the information obtained, regardless of the limited resources and other constraints. *LangDoc*[1] is then a system designed to streamline the recording and analysing process of the language data whilst mitigating errors associated with human involvement in collaborative projects, specifically for such languages including but not limited to which lacking conventionalised writing systems. Its functionality extends to both online and offline environments, making it particularly well-suited for language documentation conducted in remote locations.

In particular, this paper presents its idea, as well as system design, functionalities and features. It will also discuss the system's current limitations and outline the possible direction for future development. To illustrate the functionalities, this paper demonstrates LangDoc with a real-world use case by its application in documenting the Moklen in the Southern Thailand.

This paper makes several key contributions to the field of language documentation. Firstly, it aims to address common challenges faced in this domain, such as managing data from multiple sources, logistical difficulties in collaborative teamwork, and also extending to tackle such external limitations as the well-being of language informants. Secondly, the paper proposes features to mitigate common errors and enhance the efficacy, whilst acknowledging the essential role of trained linguists. Thirdly, the paper presents offline synchronisation feature is crucial for fieldwork in remote locations. The system allows users to collect data without an internet connection and syncs automatically when connectivity is restored. Additionally, the system's architecture allows for future integration with tools for deeper linguistic analysis to further expand its capabilities.

[1]The online system can be found at https://langdoc.piyapath.uk. For the offline programme and any other inquiries, feel free to contact the author.

Proceedings of the 3rd Workshop on NLP Applications to Field Linguistics (Field Matters 2024), pages 28–36

2 Background Issues and Related Work

Collecting vocabulary data for endangered languages presents significant challenges, particularly when the documentation effort is led by community outsiders. The conventional approach of conducting interviews and elicitation sessions with native speaker informants can be inefficient, costly, and potentially detrimental to the well-being of elderly informants who often serve as the primary sources of linguistic knowledge.

One of the major issues is the limited productivity of data collection sessions, especially with elderly informants who may have physical limitations. As observed in Moklen fieldwork, interviews with elders typically yield a maximum of 60 words per session, with several breaks required within a three-hour period. Completing a modest vocabulary list of 250 words can take at least five days of work, and more extensive projects naturally require even greater resource investment.

Another challenge arises when multiple researchers are involved in the documentation effort. Dividing informant interviews amongst project members can lead to wasted effort due to duplicate recordings of common vocabulary and the potential to miss more specific, culturally-related terms known to certain informants; not to mention the additional time to be spent merging data and identifying missing entries.

Furthermore, inconsistencies in the interpretations by different researchers can arise, especially when dealing with a semantically complex spoken languages like Moklen. To resolve these discrepancies often requires revisiting informants in person, hindering the overall progress. Even if larger team appear to be bring a faster data collection, the unique challenges of endangered language documentation suggest that a more focused approach tailored to the needs of the specific community is crucial. Overwhelming elderly informants with lots of people can lead to shorter, less productive sessions due to factors such as fatigue and discomfort.

In recent years, there have been efforts to integrate technologies for recording, transcribing, and analysing language records (Rice and Thieberger, 2018), as well as other NLP tasks (Moeller et al., 2024; Serikov et al., 2023) to language documentation. Nevertheless, most works focus on how can the data can be used to represent linguistic phenomena; little attention, however, has been given to tackle the fundamental problem of how linguists or researchers can actually and effectively collect and prepare the necessary linguistic data in the first place, especially for endangered languages with rapidly dwindling speaker populations. Of course, good tools and applications have emerged to aid in field linguistics, such as Aikuma (Bird et al., 2014), FLEx (Zook, 2024), and ELAN (Max Planck Institute for Psycholinguistics, 2023), yet often operate in silos and do not comprehensively address the multifaceted challenges faced by linguists in the field. There is a need for solutions that holistically address the data collection process whilst considering the unique logistical, ethical and community-related challenges faced when documenting such endangered languages.

The issues highlighted above point to a primary use case that the proposed system aims to address, comprising a team of field linguists with varying experience working to document the vocabulary of an endangered language spoken by a remote community with few population of elderly native speakers. In the scenario when the opportunities to work with remaining fluent speakers are increasingly limited, efficiently and sensitively collecting high-quality data are paramount.

3 The LangDoc System

LangDoc is a comprehensive system designed to streamline the language documentation process, particularly for endangered languages lacking standardised writing systems. It incorporates several key features to address the challenges identified in the background section.

3.1 Wordlist-driven Recording System

Although wordlist-driven recording is a standard practice in language documentation, LangDoc introduces significant improvements where users have their flexibility to create and customise the wordlist-based project and propose the structured workflow that minimise the complexity of working process. Unlike existing tools, LangDoc's design ensures that all entries are systematically reviewed and verified, which is particularly important in the context of endangered languages with limited speaker populations.

3.1.1 Wordlist Management

LangDoc provides a comprehensive wordlist management interface that allows users to create, edit, and organise wordlists within their projects (cf. subsubsection 3.2.1).

3.1.2 Entry Management

Each wordlist consists of individual word entries, stored in the `entry` table of the database. This table maintains information about each entry, such field as the word form, its part of speech (POS), definition, category, and its working status.

Users can add new entries to a wordlist by filling out a form that captures fields such as headword, POS, category, and meaning. Not every field is required, as users can customise the fields according to their needs. The reason for this is to accommodate various use cases of specific projects, such as creating a dictionary for the community.

3.1.3 Data Collection

The wordlist-driven recording system provides a structured and organised approach to word collection by presenting users with a list of wordlists and their associated entries. Users with the `collector` role can access the *To Collect* section, which displays wordlists that have unrecorded entries.

For each wordlist, collectors can view the percentage of entries that have been recorded, providing an overview of the progress made. By clicking on a wordlist, users are redirected to a dedicated page where they can record pronunciation data (IPA) for each entry.

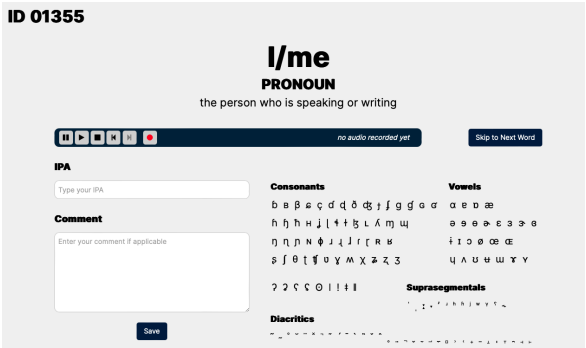

Figure 1: Sample recording interface

The data collection interface, as in Figure 1, presents the word entries sequentially, allowing users to input the IPA transcription using a character picker, add comments or notes, and navigate between entries within a wordlist. The system, however, prioritises audio recordings of words. Specifically, collectors can only record audio for each entry without providing IPA or notes. This approach helps mitigate potential biases compared to hasty transcriptions by collectors. Users also have the option to skip entries or mark them for review.

3.1.4 Instant Word Collection

In addition to the wordlist-driven approach, LangDoc offers an "Instant Word Collection" feature that enables users to quickly gather words from informants without associating them with specific wordlists in the project.

The interface is similar to normal word collection in that it allows users to record information about particular words. However, this feature also gives users more flexibility, including to either select existing informants or add new ones, to enter the head word or morpheme, and then to record the word, along with optional IPA transcription and comments.

3.2 Project Management Tools

LangDoc also provides robust project management tools, allowing users to create new projects, assign project members with specific roles (i.e. `admin`, `collector`, `analyser`), and manage project settings and preferences.

3.2.1 Project Creation

The project creation process in LangDoc is designed to be straightforward. Users can initiate the creation of a new project by providing essential information such as the project name, affiliation, and the language under study. An autocomplete feature assists users in selecting the language by suggesting matching language names or ISO 639-3 codes (International Organization for Standardization, 2007) as they type.

Once the basic project information is provided, users can choose to associate one of the existing wordlists, as shown in Table 1, with the project. Prior to the modification to include semantic category and meaning for dictionary representation, those predefined wordlists below only offer headwords and their part of speech. The other way is to proceed without a wordlist, as the customised lists can be later imported as CSV or XLSX to the project. This flexibility allows users to tailor the project setup according to their specific requirements.

After selecting the suitable wordlist, the project creator can also add people whose the LangDoc account exists within the current database to the creating project. By and large, all project settings and preferences aside from the basic information are optional and can be altered afterward.

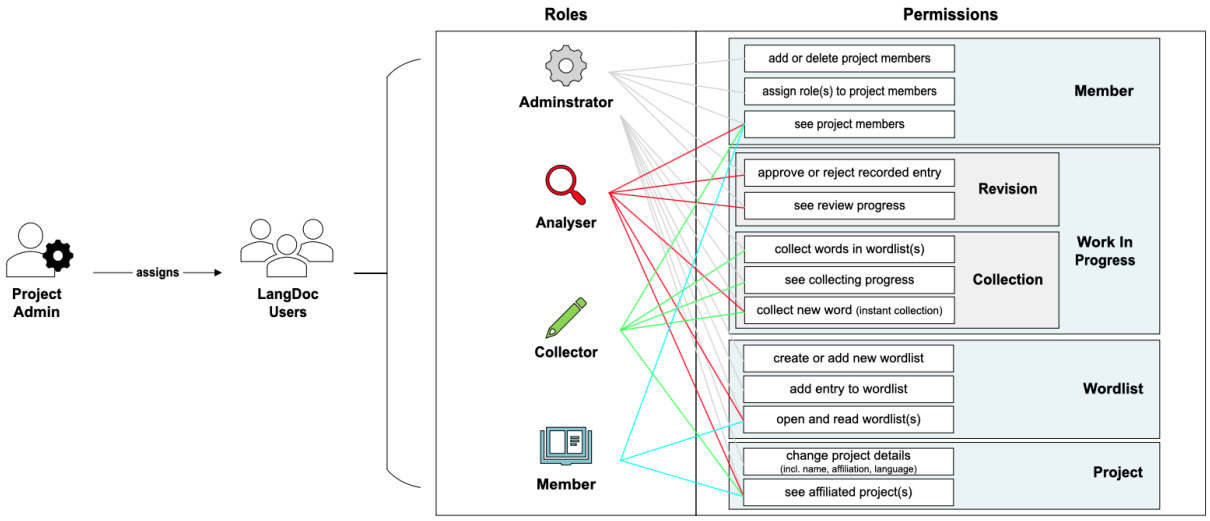

Figure 2: An RBAC diagram showing roles within a project in the LangDoc system

Wordlist	Citation
Swadesh 100	Swadesh (1971)
Swadesh 207	Swadesh (1952)
ASJP 40	Wichmann et al. (2007)
Swadesh-Yakhontov 35	Starostin (1991)
Dolgopolsky 15	Dolgopolsky (1964, 1986)
CALMSEA	Matisoff (1978)
NGSL 1.2	Browne et al. (2023)
Sign Language	Emmorey and Lane (2000)

Table 1: Predefined wordlists available in LangDoc

3.2.2 Project Assignment

LangDoc applies a role-based access control system (RBAC) to manage project members and their permissions. The project creator is automatically assigned the administrator role, which allows assigning roles with specific access levels to other members. Each user can have multiple roles within a project and roles can vary across different projects.

As illustrated in Figure 2, the available roles within a project include:

- **Project Admin**: Administrators have full control over the project, including managing members, data analysis, and data storage.
- **Analyser**: Members assigned the `analyser` role are responsible for reviewing and analysing the collected linguistic data to determine its usability and accuracy.
- **Collector**: The `collector` role involves recording and managing the collected linguistic data within the project.
- **Member**: General members have limited access and are participants in the project with standard privileges.

By assigning specific roles, LangDoc secures that the right individuals have the necessary permissions to perform their designated tasks, maintaining data security and efficient project management.

3.2.3 Project Management

LangDoc provides a dedicated project management interface that allows administrators to oversee and manage various aspects of their projects. This interface includes:

- **Project Settings**: Administrators can access and modify project preferences, including general project details and other customisation options.
- **Wordlist Management**: Administrators can create new wordlists and add entries to existing wordlists for the whole project.
- **Member Management**: Administrators can add or remove project members, as well as modify their assigned roles within the project.
- **Progress Tracking**: The project management interface provides an overview of the progress made on each wordlist, displaying the percentage of entries that have been recorded or require revision.

Through these comprehensive project management tools, LangDoc allows administrators to effectively coordinate and oversee linguistic data collection and analysis projects for the organised working environment. Whilst this section is dedicated for project administrators, some discussed functionalities can overlap across roles as seen in Figure 2.

3.3 Collaborative Review System

To enhance the quality and accuracy of the data, LangDoc includes a collaborative review system that allows senior members designated as `analysers` to collectively review, verify and refine the recorded data. Their primary tasks include listening to recorded pronunciations, verifying transcription accuracy, and making necessary corrections or annotations, so as to maintain the integrity and accuracy of the linguistic data that meet the research objectives.

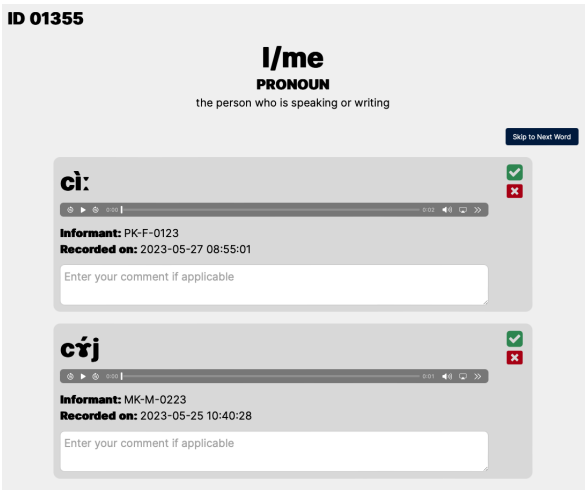

Figure 3: Sample review interface (1SG in Moklen)

The review interface in LangDoc is intuitively designed to facilitate an efficient review process. It presents all the data entries that require verification in a listed format, allowing analysers to easily navigate through them. Each entry includes detailed information such as the word, its phonetic transcription, and any notes or comments added by the collector.

Analysers can play audio recordings directly within the interface and compare them against the provided transcriptions. If discrepancies or errors are found, analysers can edit the transcriptions directly in the interface. They also have the option to add detailed comments to provide context or justification for the changes they make.

3.3.1 Collaborative Features

To promote collaboration, LangDoc includes several features that support real-time communication and data sharing amongst analysers, aside from the automatic status tagging system:

- **Commenting System**: Analysers can leave comments visible to all members on each entry to discuss discrepancies, suggest alternatives, or provide insights.
- **Change Tracking**: The system keeps a log of all changes made to each entry, including who made the change and when, to maintain transparency and accountability in the process.
- **Consensus Building**: For entries that require further discussion, analysers can flag them for review to ultimately build consensus on the most accurate transcription as the final decision.

3.4 Data Transfer

Another critical feature of the LangDoc system is its comprehensive data transfer functionality. This feature is provided due to the fact that LangDoc is designed as a tool, rather than a closed platform, to address the diverse needs of linguistic researchers and project teams for their recorded language data. It allows them to use their available data in the system, and to access and utilise their data outside the LangDoc environment.

Apart from its import functionality discussed in subsubsection 3.2.1 to serve users who are more familiar with data in other formats, The LangDoc system allows users to have complete access to their project's information via the export of various types of data, including recorded wordlists, audio recordings, and relevant metadata. Users initiate the export process by selecting the specific project or wordlist they wish to export. This ranges from the selection of specific wordlists to the entire project data. It also supports multiple export formats (i.e. CSV, JSON, XML, or ZIP files for the export includes audio recordings) for varying compatibility with various analysis tools and software.

3.5 Offline and Remote Accessibility

Field linguistics often requires researchers to work in remote areas where internet infrastructure is lacking or entirely absent. In such environments, the reliance on a constant internet connection for data collection and analysis can severely hinder the progress of linguistic documentation efforts. Recognising this, one of the significant developments of the LangDoc system is the ability to operate effectively in both online and offline environments, which is crucial for uninterrupted linguistic data collection in remote field locations with sporadic or non-existent internet connectivity. A detailed explanation of the technical implementation, including data synchronisation, local storage, and system architecture will be presented in the following section.

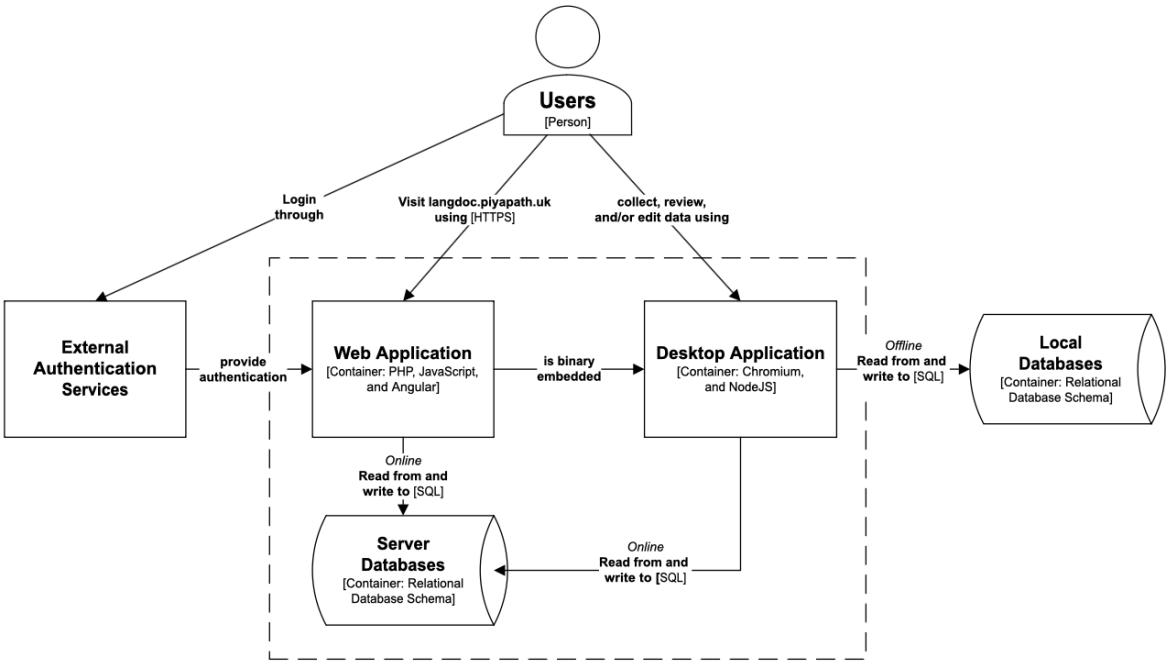

Figure 4: A high-level C4 container diagram of LangDoc system

3.6 System Architecture

As shown in Figure 4, LangDoc follows a client-server architecture with web-based user interfaces (i.e. web application and desktop application) interacting with a backend server that stores data in relational databases.

LangDoc supports external authentication methods, allowing users to authenticate using their accounts from external providers. This external authentication component communicates with the respective authentication providers' APIs to facilitate user login, registration, and account management.

The main interface for LangDoc is a web-based application. On the server-side, the application employs PHP to handle data processing, database interactions, and server-side logic, whilst Nginx web server is responsible for serving the application and handling HTTP requests. On the client-side, HTML, CSS, and JavaScript are used to create the user interface, handle user interactions, and provide a responsive and dynamic experience. The application also incorporate Angular, a JavaScript frameworks, to facilitate efficient development and maintainability. Though accessible on various devices, the interface is optimised for PC usage

LangDoc also offers the desktop interface designed specifically for offline word collection and temporary local storage, using an SQLite database to store linguistic data and project information. When the desktop application is online, it synchronises the locally stored data with the cloud database server. This process involves uploading any new or modified data to the server and downloading any updates or changes made by other users or collaborators. The desktop application is built using ChromiumOS rendering and Node.js for cross-platform compatibility, which allows for the creation of desktop applications using web technologies like HTML, CSS, and JavaScript to create a consistent user experience across systems.

The data synchronisation between the offline desktop application and the server is a crucial aspect of the LangDoc system. The mechanism adopts long-polling protocols to establish a connection between the desktop application and the cloud server. The desktop application stores data locally, keeping track of any new, modified, or deleted entries using timestamps during offline. It initiates the synchronisation process when detecting an internet connection. Timestamping is employed to prevent conflicts and determine which changes should take precedence, as the system allows multiple entries supported by the review system.

The deployment architecture of the LangDoc system varies depending on specific requirements and infrastructure available. For local development and testing purposes, the system is deployed on a virtual environment, with the web application run-

ning on Apache and the database server running on the same machine. For staging or production, the system is implemented on AWS cloud platform, hosted on an Canonical Ubuntu 22.04 E2 instance. The deployment architecture incorporates load balancing, caching, and optimisation techniques for scalability and availability. For security measures, SSL/TLS encryption and firewalls are implemented to protect the system and user data.

4 Evaluation and Case Study

It is always difficult to find a good matrix to measure the performance of software development systems such as LangDoc. However, evaluating its impact on language documentation projects is crucial for understanding its effectiveness and efficiency.

4.1 Comparative Analysis

To quantitatively evaluate the performance of LangDoc against traditional paper-based methods and computer-assisted audio recording, an experiment was conducted involving eight non-Vietnamese participants collecting Vietnamese vocabulary using the Swadesh-Yakhontov 35 wordlist across 3 different methods. Figure 5 visualises the central tendency distribution of time taken for each methods.

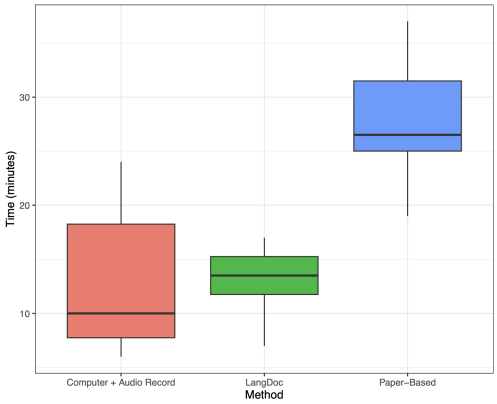

Figure 5: A boxplot of time taken for documenting tasks using different methods

One-way ANOVA showed a significant difference in mean times among methods ($F(2, 21) = 19.33$, $p < 0.001$). Additionally, a post-hoc Tukey's HSD test indicated the paper-based method ($\bar{x} = 27.88$, $s = 5.59$) took significantly longer than the computer with audio recording ($\bar{x} = 13.00$, $s = 7.03$) and LangDoc ($\bar{x} = 13.13$, $s = 3.18$) methods.

Overall, Figure 5 shows that both computer-assisted and LangDoc significantly improve data collection efficiency over paper-based methods, which, although gradually decreasing in modern fieldwork, still occur in certain scenarios. Besides, whilst traditional measure appears faster in median time, I argue that the consistency and accuracy of LangDoc's data collection process offer substantial long-term benefits by reducing the need for subsequent corrections and reverifications. Still, it is not appropriate to claim from the result as the experiment only involved the collection of 35 words and did not test the review process. Our case study on the Moklen language in the following section further demonstrates these advantages in a real-world setting.

4.2 Case Study: Documenting the Moklen Language

The case study of documenting the Moklen language in Phuket and Phang-nga, Thailand stands as evidence of LangDoc's effectiveness in addressing challenges faced by field linguists working with endangered languages and remote communities, as highlighted in the section 2.

Like many endangered languages, Moklen is predominantly spoken by the older generation, typically those above 50 years old who are Moklen-Thai bilingual (Pittayaporn and Choemprayong, Forthcoming). However, fluent speakers of the language are mostly amongst those exceeding 70 years old, restricting potential informants to only the elderly population. Despite their willingness to help teach the language and share knowledge, the documentation process itself can present unforeseen challenges due to the physical limitations that often come with age. Unlike younger generations having more stamina, extended recording sessions usually require elders to remain seated for longer periods. They may also need to repeat information or clarify pronunciations, which can be tiring. Additionally, the nature of documentation, where the duration are unstructured and depend on the flow of the conversation, is likely to inadvertently cause discomfort for elderly informants.

LangDoc's workflow and offline capabilities allowed researchers to conduct sessions at a comfortable pace for the elderly informants, reducing the chance of discomfort. The system facilitated the data collection process by preventing the clustering of records for words already documented. This feature not only accelerated the overall collection process but also minimised unnecessary post-processing tasks.

The ability to work offline and synchronise data

later proved invaluable, enabling researchers to focus on building rapport with informants. This led to more productive sessions and richer linguistic data collection. The motivation behind integrating offline functionality into LangDoc stems from the need to support fieldwork research in any setting, particularly for documenting endangered languages spoken by isolated communities like Moklen in Ko Phra Thong Island, Thailand. The offline capabilities allow greater flexibility, enabling researchers to collect data in the field and later synchronise it with central servers when internet access is restored, integrating into the broader project database.

Moreover, LangDoc's flexibility in both environment and wordlist configuration allowed the Moklen project to recollect sample audio recordings for each lemma, facilitating the production of a comprehensive Moklen dictionary. As of now, the Moklen language documentation project expects to compile a comprehensive database of over 1,000 words, complete with audio recordings, IPA transcriptions, and cultural annotations.

The success of the Moklen language documentation project underscores LangDoc's value in enhancing the efficacy and effectiveness of language documentation efforts, particularly in challenging field conditions. The system's ability to address the unique needs of endangered language communities and remote locations highlights its potential to support the documentation of linguistic diversity worldwide, preserving invaluable cultural heritage for future generations.

5 Discussion and Future Directions

The LangDoc system represents a step forward in optimising language documentation process, particularly for endangered languages in remote communities. However, it is essential to acknowledge the limitations of the current system. Whilst excelling in data collection and organisation, LangDoc primarily focuses on the preliminary stages of language documentation, currently limited to managing wordlists, transcriptions, and basic metadata. Additionally, the system's reliance on manual input and human involvement, even if mitigated through its collaborative features, may still introduce potential biases or inconsistencies, particularly in the transcription and annotation processes.

Integrating LangDoc with state-of-the-art NLP techniques could significantly enhance its capabilities to help linguists doing their works. For example, automated transcription and annotation tools could reduce manual effort and potential biases from humans, allowing linguists to provide essential oversight, control quality, and go further with analysis of complex linguistic phenomena beyond lexicons. Additionally, incorporating machine learning models trained on the collected data could assist in developing low-resource technologies, such as machine translation, parsing, and ASR systems for the documented languages.

Exploring ways to involve language communities more actively in the documentation process could foster a sense of ownership and promote the preservation of linguistic heritage. This could involve developing user-friendly interfaces for community members to contribute to data collection, validation, and dissemination efforts, in addition to the tool used solely by linguists.

6 Conclusion

This paper presented LangDoc as a system to address challenges in documenting endangered languages without standardised writing not only in the form of software tools but also via presenting logical steps for human workflow. By incorporating project management, wordlist-driven recording, collaborative review, and offline access, it improves documentation efficiency and quality. The Moklen case study demonstrated LangDoc's capabilities in tackling data duplication, verification bottlenecks, and accommodating elder informants. Whilst not a panacea, LangDoc streamlines workflows and enhances collaborative project effectiveness, helping preserve linguistic diversity and sustain endangered languages in its most foundational process.

Acknowledgments

The work would not be possible with out the help and participation of the Moklen community. I sincerely thanks Dr Pittayawat Pittayaporn for the opportunity to participate in his Moklen fieldwork (and this workshop deadline wake-up call). My gratitude extends to Dr Songphan Choemprayong for his endless support and critical comment on the near-final version. I also appreciate to my colleagues, Theera-anuchit Chalapinyo and Nichanan Pornsirivorarak, for their patience during the project's pilot period. Finally, I am grateful to three anonymous reviewers for their attentive reading of the paper and useful comments. Needless to say, only I am responsible for any remaining errors.

References

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 1–5, Baltimore, Maryland, USA. Association for Computational Linguistics.

H. Andrew Black and Gary F. Simons. 2006. The SIL FieldWorks Language Explorer approach to morphological parsing. In *Proceedings of the 10th annual Texas Linguistics Society conference: Computational Linguistics for Less-Studied Languages*, pages 37–55, Austin, Texas, USA.

C. Browne, B. Culligan, , and J. Phillips. 2023. New general service list 1.2.

Bradley Wendell Compton. 2014. Ontology in information studies: without, within, and withal knowledge management. *Journal of Documentation*, 70:425–442.

Aharon B. Dolgopolsky. 1964. Gipoteza drevnejšego rodstva jazykovych semej severnoj evrazii s verojatnostej točky zrenija [a probabilistic hypothesis concering the oldest relationships among the language families of northern eurasia]. *Voprosy Jazykoznanija*, 2:53–63.

Aharon B. Dolgopolsky. 1986. A probabilistic hypothesis concerning the oldest relationships among the language families of northern eurasia. In Vitalij V. Shevoroshkin, editor, *Typology, relationship and time. A collection of papers on language change and relationship by Soviet linguists*, pages 27–50. Karoma Publisher, Ann Arbor. Originally published in 1964 as "Gipoteza drevnejšego rodstva jazykovych semej Severnoj Evrazii s verojatnostej točky zrenija" and translated from the Russian by V. V. Shevoroshkin.

Joel Robert William Dunham. 2014. *The online linguistic database: software for linguistic fieldwork*. Ph.D. thesis, The University of British Columbia.

Karen Emmorey and Harlan L. Lane. 2000. *The Signs of Language Revisited: An Anthology To Honor Ursula Bellugi and Edward Klima*. Psychology Press.

International Organization for Standardization. 2007. ISO 639-3:2007, Codes for the representation of names of languages – Part 3: Alpha-3 code for comprehensive coverage of languages.

James A. Matisoff. 1978. *Variational Semantics in Tibeto-Burman*. Institute for the Study of Human Issues.

Max Planck Institute for Psycholinguistics. 2023. ELAN (Version 6.7) [Computer software].

Sarah Moeller, Godfred Agyapong, Antti Arppe, Aditi Chaudhary, Shruti Rijhwani, Christopher Cox, Ryan Henke, Alexis Palmer, Daisy Rosenblum, and Lane Schwartz, editors. 2024. *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, St. Julians, Malta.

Pittayawat Pittayaporn and Songphan Choemprayong. Forthcoming. A proposal for a thai-based moklen orthography. *Language Documentation and Conservation*.

Jens Rasmussen and Kim J. Vicente. 1989. Coping with human errors through system design: implications for ecological interface design. *International Journal of Man-Machine Studies*, 31(5):517–534.

Keren Rice and Nicholas Thieberger. 2018. 225Tools and Technology for Language Documentation and Revitalization. In *The Oxford Handbook of Endangered Languages*. Oxford University Press.

Oleg Serikov, Ekaterina Voloshina, Anna Postnikova, Elena Klyachko, Ekaterina Vylomova, Tatiana Shavrina, Eric Le Ferrand, Valentin Malykh, Francis Tyers, Timofey Arkhangelskiy, and Vladislav Mikhailov, editors. 2023. *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia.

Sergej Starostin. 1991. *Altajskaja problema i proischoždenije japonskogo jazyka [The Altaic problem and the origin of the Japanese language]*. Nauka, Moscow.

Morris Swadesh. 1952. Lexico-statistic dating of prehistoric ethnic contacts. with special reference to north american indians and eskimos. *Proceedings of the American Philosophical Society*, 96(4):452–463.

Morris Swadesh. 1971. *The origin and diversification of language: Edited post mortem by Joel Sherzer*. Aldine, Chicago.

Daan van Esch, Ben Foley, and Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 14–22, Honolulu. Association for Computational Linguistics.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2007. Guidelines for preparing 40-word lists for languages to be included in the ajsp database. *The ASJP Database*.

Penwipa Yooyen. 2013. *Tone variation of Thai Song by age group in Ratchaburi province*. Ph.D. thesis, Mahidol University.

Ken Zook. 2024. *FLEx 9.1 Conceptual Model*. SIL International.