Most of the world's printed texts published before the widespread adoption of word processing software are not available in digital, machine-readable format [1,2]. This is particularly true for texts of Indigenous languages collected by Europeans during early contact and later by anthropologists, including grammars, lexica, parallel texts, religious documents, and community newspapers. Producing machine-readable versions of these texts is a crucial step in reclaiming cultural and linguistic materials for the communities who speak these languages and the researchers with whom they choose to collaborate.

Optical character recognition (OCR), the technology that converts images of text into machine-readable text files, is a mature technology for Latin alphabet data, and off-the-shelf OCR toolkits yield strong accuracy for high-resource languages like English. This does not guarantee, of course, high accuracy for other languages, particularly those written with characters and diacritics that are different from those that appear in the primarily modern European texts used to train these models.

Some popular OCR toolkits [3] can adapt or "fine-tune" pre-trained models to new data which helps alleviate these problems for under-resourced languages, but these toolkits are not ideal for endangered languages [4] and are designed for experienced programmers with access to robust computing resources [5]. The CMU Linguistic Annotation Backend (CMULAB [6]) offers an alternative with a simple interface that anyone can use to fine-tune an existing OCR model to a new language using hand-corrected OCR output.

We explore the utility of the CMULAB OCR interface for accelerating the digitization of 17th-century printed texts in the Formosan language Siraya. The Formosan languages are members of the larger Austronesian language family, which also includes Tagalog and Hawaiian. Specifically we ask: (1) Can we use CMULAB to digitize archival texts with non-standard characters and diacritics? (2) Can accuracy be improved with fine-tuning? (3) Is correcting OCR output faster than transcribing from scratch? (4) Does transcription speed improve as OCR accuracy increases?

We conclude that CMULAB is a very effective and easy-to-use tool for converting scanned images of archival text into machine-readable text. Without fine-tuning, we find an average character error rate of 5.7%, which is reduced to 3.6% with fine-tuning on hand-corrected output. Transcribing by correcting OCR was dramatically faster for our 5 transcribers, reducing transcription time by 52% on average. Correcting output from a fine-tuned model reduces time to transcribe by an additional 27%, highlighting the value of adapting models to language-specific data.

References

[1] Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, Google Books Team, Joseph P. Pickett et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331:6014, 176-182.

[2] Nguyen, Thi Tuyet Hai, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-OCR processing approaches. ACM Computing Surveys (CSUR) 54:6, 1-37.

[3] https://github.com/tesseract-ocr/tesseract

[4] Schwartz, Lane, Emily Chen, Hyunji Hayley Park, Edward Jahn, and Sylvia LR Schreiner. 2021. A Digital Corpus of St. Lawrence Island Yupik. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 31-40.

[5] Agarwal, Milind, and Antonios Anastasopoulos. 2024. A Concise Survey of OCR for Low-Resource Languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, 88-102.

[6] Sheikh, Zaid, Antonios Anastasopoulos, Shruti Rijhwani, Lindia Tjuatja, Robbie Jimerson, and Graham Neubig. 2024. CMULAB: An Open-Source Framework for Training and Deployment of Natural Language Processing Models. arXiv preprint arXiv:2404.02408.