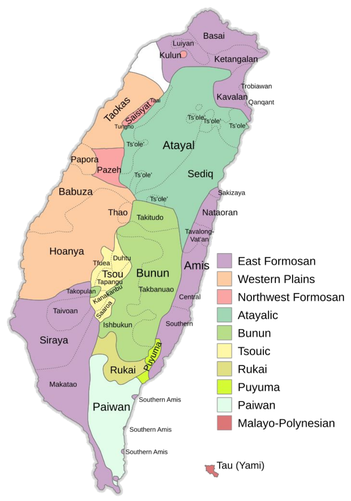
Reclaiming Archival Texts with User-Friendly OCR

Omar Tall[1], Éric Le Ferrand[1], Hunter Scheppat[1], Joshua Hartshorne[2], and Emily Prud'hommeaux[1]

[1]Boston College [2]MGH Institute of Health Professions

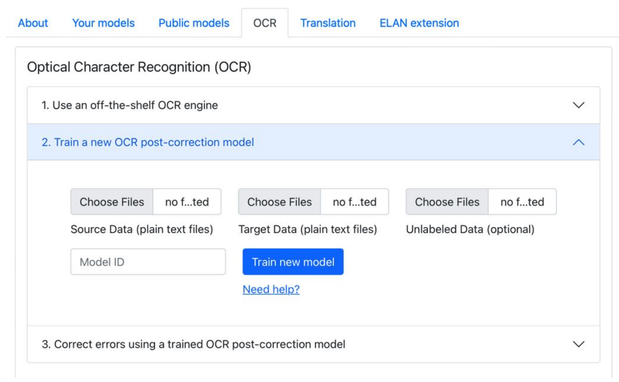
BACKGROUND

- Most pre-20thC non-European texts not yet digitized. [1]
- Especially true for Indigenous language data collected during early contact: grammars, lexica, texts, translations.
- Producing machine-readable versions of these texts is crucial for reclaiming linguistic and cultural knowledge.
- OCR can help, but few OCR models are trained on Indigenous language data from that era, performance is weak. [2]
- How can we leverage existing OCR tools for this task?



DATA

- 16 Formosan languages: Austronesian, Indigenous to Taiwan.
- Focus: **Siraya**, no longer spoken natively but being revitalized.
- 17th century Bible translations produced by Dutch missionaries.
- Not yet in machine-readable format.
- Includes unusual diacritics, characters.



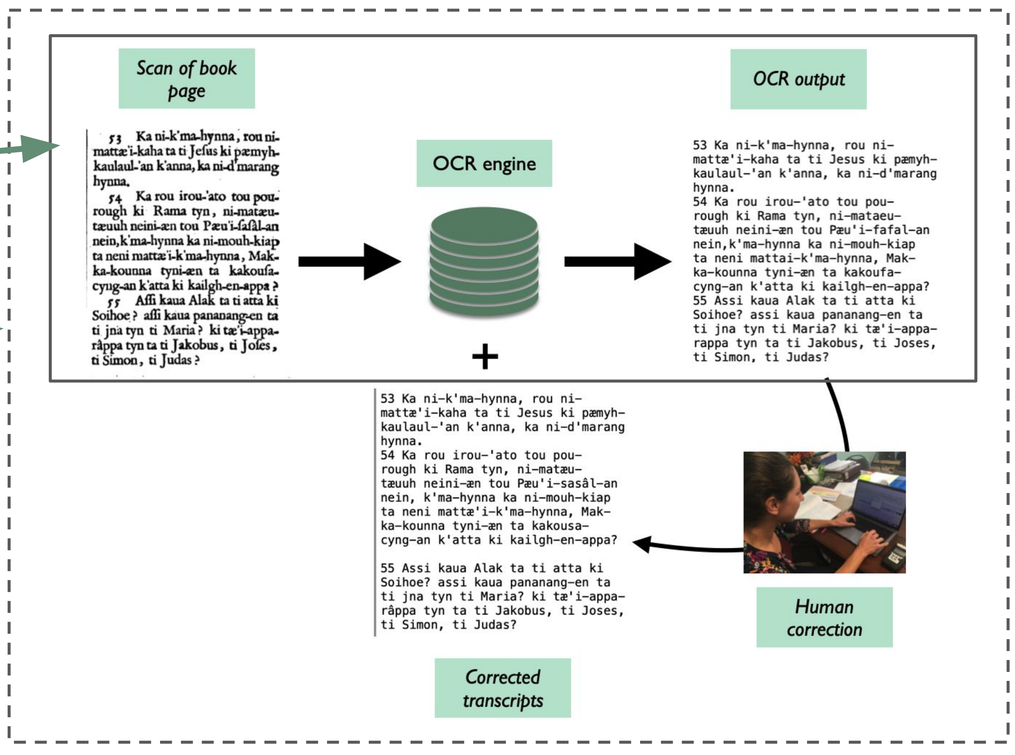
- CMULAB [3] web interface, front-end for Google Cloud Vision OCR (CNN+LSTM)
- Strong performance on Latin scripts.
- Allows tuning on corrected output.
- Very user-friendly interface.

METHOD

One "off the shelf" condition:
Output from GCV OCR with no tuning

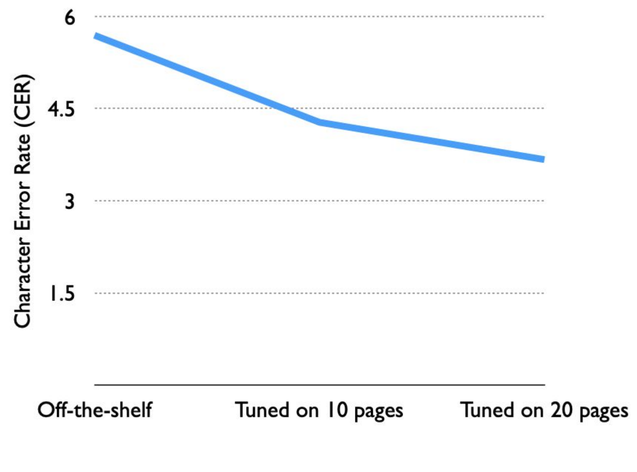
Two tuning conditions:
Output when tuned on 10 or 20 hand-corrected pages

Evaluation:
CER for 3 conditions
Time-to-transcribe for 5 transcribers from scratch and when correcting OCR



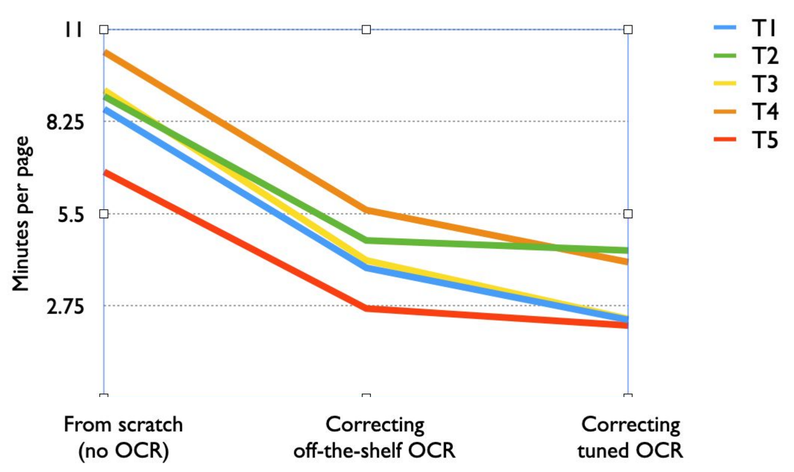
RESULTS

CER off-the-shelf and with tuning on hand-corrected output



- ✓ CER drops from 5.7 off-the-shelf to 3.6 with tuning on 20 pages.
- ✓ Time-to-transcribe reduced dramatically when correcting OCR output (average 52%).
- ✓ Further reduced when using lower-CER fine-tuned output.

Time-to-transcribe from scratch and when correcting OCR output



CONCLUSIONS

- OCR tools off-the-shelf provide remarkably accurate results even for archival texts with unusual characters.
- Results can be further improved via tuning.
- Lower CER leads to improved transcription times.

FUTURE WORK

- Existing OCR tools perform poorly on glossed text, translations in a different script (e.g., Mandarin).
- We plan to develop bespoke system for identifying glossed/translated examples in scanned texts.

ACKNOWLEDGEMENTS

This material is based upon work supported in part by the NSF Grant No. 2319296. We are grateful for the support of the ILRDF and our other collaborators in Taiwan.

REFERENCES

[1] Michel et al. 2011. Quantitative analysis of culture using millions of digitized books. Science 331:6014, 176-182. [2] Schwartz et al. 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. [3] Sheikh et al.. 2024. CMULAB: An Open-Source Framework for Training and Deployment of Natural Language Processing Models. arXiv preprint arXiv:2404.02408.