# Working with Optical Character Recognition to Document an Understudied Language: Challenges and Opportunities

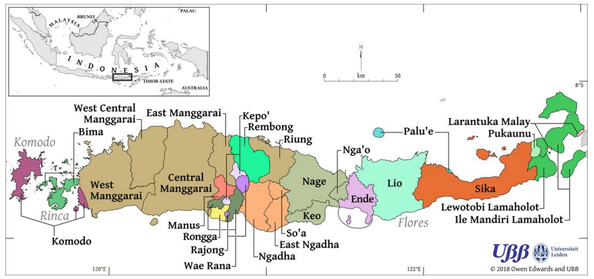
Erin McCartney[1], Grace B. Wivell[1], Fransiskus X. Mbete[2]
[1]Stony Brook University and [2]Wolondopo, Indonesia

## Introduction

- Field linguists, researchers of underdocumented languages, and community members can benefit from technology that assists in quickly transcribing written language data.
- Most elicited data is spoken, but written data can reveal different uses of the language.
- Like how audio data can be transcribed with Automatic Speech Recognition (ASR), written data can be transcribed into searchable, processable text using **Optical Character Recognition (OCR)**.
- However, both of these technologies are less accessible for underdocumented languages and their communities that need them.

## Background: Lio Language

- **Lio** is an Austronesian language spoken in Central Flores, Indonesia with 220,000 speakers (Eberhard 2020).
- The **2019 Linguistic Fieldwork and Documentation Training in Indonesia Program** collected data from the community, which is available on PAcific and Regional Archive for DIgital Sources in Endangered Cultures (**PARADISEC**).
- Most of this data comes from recorded speech (conversations and monologues), but also present in PARADISEC are photo images of pages of a book written in Lio- a religious text called a missalette containing common prayers and a guide to the Catholic mass.



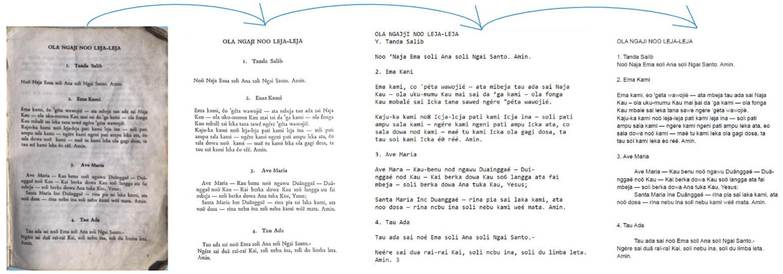

*Lio Region of Flores, Indonesia. Map provided by Unit Bahasa dan Budaya (Language and Culture Unit) Kupang, Indonesia*

## Background: Optical Character Recognition

- Tesseract OCR, a free technology for automatically transcribing text from images.
- Can recognize over 100 languages, but Lio is not one of them (Google 2008).

## Method 1: Manually Correcting Errors

- Lacking familiarity with the language's alphabet, lexicon, etc., Tesseract makes errors based on whatever language setting is being used.
- Cleaning up the image using a software such as ImageMagick (The ImageMagick Development Team 2021) helps reduce errors, but not enough.



1. Original photographed image of physical page
2. Processed image: black and white, noise removal, deskew *(using ImageMagick's textcleaner script)*
3. Automated transcription using Tesseract Optical Character Recognition technology *(can be improved by training on Lio data)*
4. Manually proofread and edited final transcription

## Method 2: Training on Lio Data

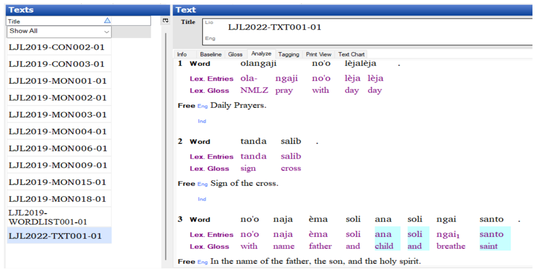
- Tesseract provides the option to train the OCR to recognize a new language based on data you provide including example images and word lists.
- Two versions with different methods: Modern versions (4 and 5) use neural nets and require weeks to train, large amounts of data, and a powerful computer; Legacy versions (1-3) are simpler and faster but less accurate and there are fewer information resources for troubleshooting.

| | English | French | German | Indonesian | Lio |
|---|---|---|---|---|---|
| Original | 65.29% | 66.32% | 65.58% | 65.73% | TBD |
| Cleaned | 7.83% | 5.76% | 7.09% | 7.24% | TBD |
| Ideal | 4.43% | **1.62%** | 3.84% | 5.47% | TBD |

*Comparison of Character Error Rates (Insertions + Deletions + Substitutions / Characters in Reference) of different language settings and image qualities.*

Note: The errors for the original image are mostly due to deletions, which have more to do with the quality of the image and rotation of the text than the language setting, so I predict that Lio will still have a significant error rate. For the cleaned image and recreated/ideal image, deletions are rare and most errors are substitutions, so I predict that Lio will have very good results- maybe even close to 0% errors with the ideal image.

## Results so Far




*An example of a transcribed page successfully entered into FLEX. Known words are automatically glossed and new entries can be made for unknown words.*

Note: Some unknown words can be guessed from context, especially if they're very common in the text (ex: "soli"). The texts are also easier to translate due to their content being culturally familiar (Catholic prayers). We come up with hypotheses for unknown words and then consult a community member about the meaning and usage.

## Unique Factors & Challenges

- Lio being an analytic language with little morphology (Elias 2018) means that there's a low word to lemma ratio, making the word list files used by Tesseract for training (taken from our database's dictionary) more useful.
- Differences between informal/spoken language and formal/written language in a religious context means that many words aren't in our database already. The word list file needs to be regularly updated for retraining.
- In order to upload to Fieldworks Language EXplorer (FLEX), the text needs to be converted into a compatible format (all lowercase, space before punctuation).
- The orthography used in our database (and currently used by the community) is slightly different from that used in the missalette.
  é → e, e → è; VV̈ (ex: oö) → V'V (o'o); 'b, 'd, 'g → bh, dh, gh
- The village of Wolondopo lacks scanner technology, so the pictures had to be taken on a cellphone. The book is also very old and fragile. These two factors caused the image quality to be low.

## Selected References

Eberhard, D., Simons, G., Fennig, C. (eds.). 2020. Ethnologue: Languages of the World. Twenty third edition. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.
Google (2008). Tesseract OCR. Retrieved from https://github.com/tesseract-ocr
- https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html
- https://tesseract-ocr.github.io/tessdoc/Home.html

The ImageMagick Development Team. (2021). ImageMagick. Retrieved from https://imagemagick.org
Elias, A. (2018). Lio and the Central Flores languages.
Yanti (collector). 2019. Recordings of various texts in Lio, a language spoken in Ende, East Nusa Tenggara, Indonesia. Collection LJL2019 at catalog.paradisec.org.au [Open access].