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

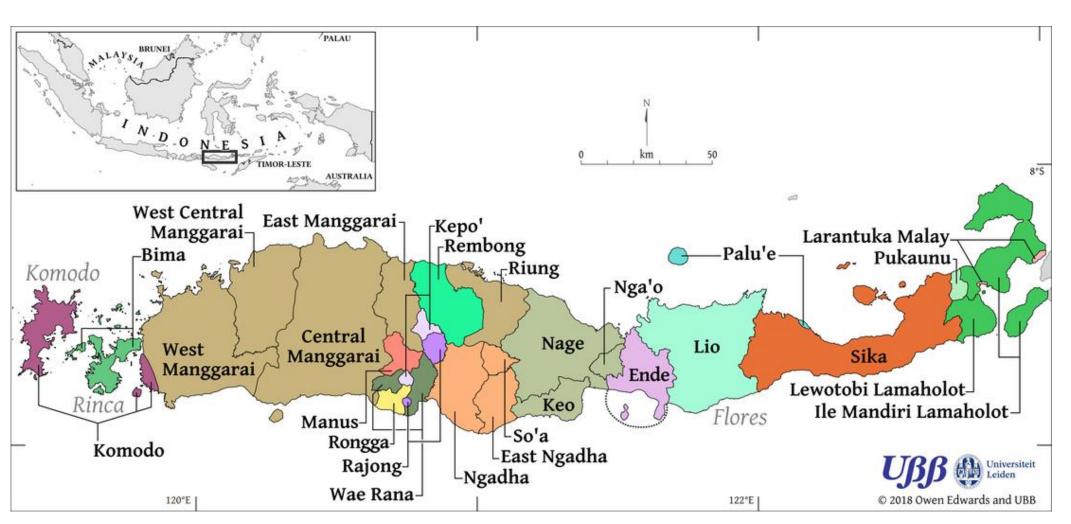


Introduction

- 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 Lioa 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).

Erin McCartney¹, Grace B. Wivell¹, Fransiskus X. Mbete² ¹Stony Brook University and ²Wolondopo, Indonesia

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.

OLA NGAJI NOO LEJA-LEJA	ola ngaji noo leja-leja	OLA NGAJjI Y. Tanda S
I. Tanda Salib	1. Tanda Salib	Noo 'Naja
entroping on April and all and a setting a compact that is a		2. Ema Kan
Noō Naja Ema soli Ana soli Ngai Santo. Amin.	Noö Naja Ema soli Ana soli Ngai Santo. Amin.	Ema kami,
a. Ema Kami	2. Ema Kami	Kau — ola Kau mobalé
Ema kami, éo 'géta wawojië — ata mbeja tau ada mi Naja Kau — ola uku-mumu Kau mai sai da 'ga kami — ola fonga Kau mbalé sai leka tana sawé ngére 'géta wawojië. Kaju-ka kami noö leja-leja pati kami leja ina — soli pati ampu sala kami — ngére kami ngeni pati ampu leka ata, éo sala dowa noö kami — maë tu kami leka ola gagi dosa, ta tau sol kami leka éo réë. Amin.	Ema kami, éo 'géta wawojië — ata mbeja tau ada sai Naja Kau — ola uku-mumu Kau mai sai da 'ga kami — ola fonga Kau mbalé sai leka tana sawé ngére 'géta wawojië. Kaju-ka kami noö leja-leja pati kami leja ina — soli pati ampu sala kami — ngére kami ngeni pati ampu leka ata, éo sala dowa noö kami — maë tu kami leka ola gagi dosa, ta tau soī kami leka éo réë. Amin.	Kaju-ka ka ampu sala sala dowa tau soi ka
 Ave Maria Ave Maria — Kau benu noö ngawu Dulnggaë — Dui- nggaë noö Kau — Kai berka dowa Kau soö langga ata fai 	3. Ave Maria	3. Ave Mar Ave Mara – nggaé nod
mbeja — soli berka dowa Ana tuka Kau, Yesua; Santa Maria Ine Duänggač — rina pia sai laka kami, ata noō dosa — rina nebu ina soli nebu kami wéë mata. Amin.	Ave Maria — Kau benu noö ngawu Dušnggaë — Duš- nggaë noö Kau — Kai berka dowa Kau soö langga ata fai mbeja — soli berka dowa Ana tuka Kau, Yesus; Santa Maria Ine Dušnggaë — rina pia sai laka kami, ata noö dosa — rina nebu ina soli nebu kami wéë mata. Amin.	mbeja — so Santa Mari noo dosa —
4 Tau Ada		4. Tau Ada
Tau ada sai noö Ema soli Ana soli Ngai Santo,- Ngére sai duā raī-raī Kai, soli nebu ina, soli du limba leta. Amin.	4. Tau Ada Tau ada sai noö Ema soli Ana soli Ngai Santo	Tau ada sa
	Ngére sai duã raī-raĩ Kai, soli nebu ina, soli du limba leta. Amin.	Neére sai Amin. 3
1. Original	2. Processed image: black	3. Aut
photographed image	and white noise removal	Tesse

deskew (using ImageMagick's textcleaner

tomated transcription using eract Optical Character Recognition technology (can be improved by training on Lio data)

Method 2: Training on Lio Data

of physical page

- 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

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.

Ema soli Ana soli Ngai Santo. Amin

NOO LEJA-LEJA

co 'péta wawojié — ata mibeja tau ada sai Naja I-mumu Kau mai sai da 'ga kami — ola fong i Icka tana sawed ngére "péta wawojié. no8 Icja-lcja pati kami Icja ina — soli pat i – ngére kami ngeni pati ampu Icka ata, co d kami — maé tu kami Icka ola gagi dosa, ta mi Icka é0 réé. Amin.

(au-benu nod ngawu Duainggaé — Dui-Kau - Kai berka dowa Kau so6 langga ata fa a Inc Duanggaé — rina pia sai laka kami, ata - rina ncbu ina soli nebu kami weé mata. Amin

noé Ema soli Ana soli Ngai Santo.a rai-rai Kai, soli ncbu ina, soli du limba leta.

I. Tanda Salib Noö Naja Ema soli Ana soli Ngai Santo. Ami

OLA NGAJI NOO LEJA-LEJA

ma kami, éo 'géta wawojië --- ata mbeja tau ada sai Naj (au – ola uku-mumu Kau mai sai da 'ga kami — ola for u mbalé sal leka tana sawé ngère 'géta wawojië. Kaiu-ka kami noö leia-leia pati kami leia ina — soli pa ampu sala kami — ngére kami ngeni pati ampu leka ata, e ala dowa noö kami — maē tu kami leka ola gagi dosa, t tau soï kami leka éo réë. Am

Ave Maria — Kau benu noö ngawu Duänggaë — Duä nggaë noö Kau — Kai berka dowa Kau soö langga ata fai anta Maria Ine Duãnggaẽ — rina pia sai laka kami, ata õ dosa — rina nebu ina soli nebu kami wéē mata. Amin

Tau ada sai noö Ema soli Ana soli Ngai Sant gére sai duä raï-raï Kai, soli nebu ina, soli du limba leta.

4. Manually proofread and edited final transcription

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

Results so Far

Title Lio L. Eng Info Baseline Gia 1 Word Lex. Entries Lex. Gloss Free Eng Daily I Ind 2 Word Lex. Entries	olanga ola- NMLZ	Tagging ji ngaji		Text Chart Ièjalè Ièja day	eja . lèja			
Info Baseline Gid 1 Word Lex. Entries Lex. Gloss Free Eng Daily I Ind 2 Word	olanga ola- NMLZ Prayers.	ji ngaji pray	no'o no'o	lèjalè lèja	eja . lèja			
1 Word Lex. Entries Lex. Gloss Free Eng Daily I Ind 2 Word	olanga ola- NMLZ Prayers.	ji ngaji pray	no'o no'o	lèjalè lèja	eja . lèja			
Lex. Gloss Free Eng Daily I Ind 2 Word	NMLZ Prayers.	pray		-				
Free Eng Daily I Ind 2 Word	Prayers.		with	day	day			
Ind 2 Word	-	salib						
2 Word	tanda	salib						
	tanda	salib						
Lex Entries								
Lex. Gloss	tanda sign	salib cross						
	2							
Ind								
3 Word	no'o	naja	èma	soli	ana	soli	ngai	santo
Lex. Entries		naja	èma	soli	ana	soli	ngai ₁	santo
Lex. Gloss	with	name	father	and	child	and	breathe	saint
	Ind 3 Word Lex. Entries	Mord no'o Lex. Entries no'o	3 Word no'o naja Lex. Entries no'o naja	Ind 3 Word no'o naja èma Lex. Entries no'o naja èma	Ind 3 Word no'o naja èma soli Lex. Entries no'o naja èma soli	Ind 3 Word no'o naja èma soli ana Lex. Entries no'o naja èma soli ana	Ind 3 Word no'o naja èma soli ana soli Lex. Entries no'o naja èma soli <mark>ana soli</mark>	Ind 3 Word no'o naja èma soli ana soli ngai Lex. Entries no'o naja èma soli <mark>ana soli</mark> ngai ₁

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.
- converted into a compatible format (all lowercase, space before punctuation). slightly different from that used in the missalette.
- In order to upload to Fieldworks Language EXplorer (FLEX), the text needs to be • The orthography used in our database (and currently used by the community) is $e \rightarrow e, e \rightarrow e$; VV (ex: oo) \rightarrow V'V (o'o); 'b, 'd, 'g \rightarrow 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

International. Online version: http://www.ethnologue.com

 https://tesseract-ocr.github.io/tessdoc/Home.html Elias, A. (2018). Lio and the Central Flores languages. LJL2019 at catalog.paradisec.org.au [Open access].



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

- Google (2008). Tesseract OCR. Retrieved from https://github.com/tesseract-ocr
 - https://tesseract-ocr.github.io/tessdoc/Data-Files-in-different-versions.html

Eberhard, D., Simons, G., Fennig, C. (eds.). 2020. Ethnologue: Languages of the World. Twenty third edition. Dallas, Texas: SIL

The ImageMagick Development Team. (2021). ImageMagick. Retrieved from https://imagemagick.org

Yanti (collector). 2019. Recordings of various texts in Lio, a language spoken in Ende, East Nusa Tenggara, Indonesia. Collection