Making level ground for future linguistic work to stand on

Blaine Billings
University of Hawai‘i at Mānoa
AIFIS-MSU Conference on Indonesian Studies
18 June 2025

Roadmap

1. Background
2. Improving research practices
3. Future linguistic work
4. Conclusion

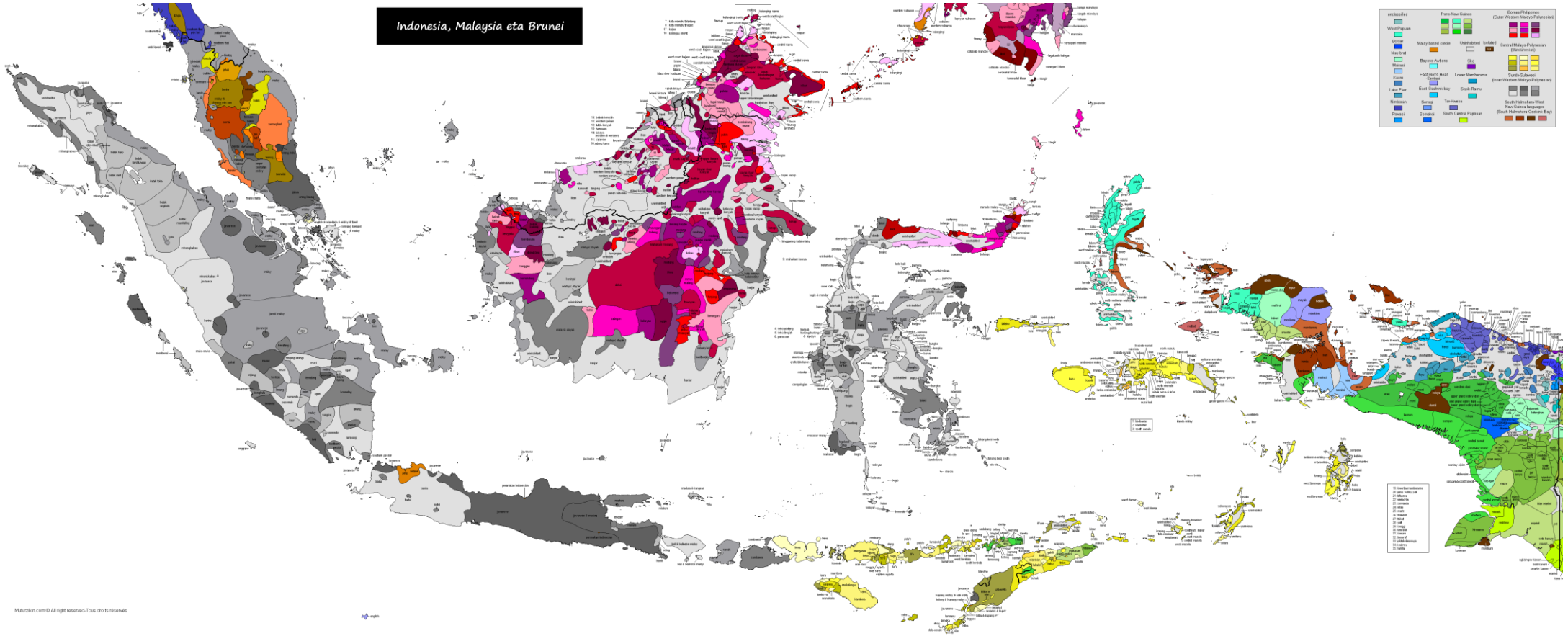

Figure 1. Languages of Indonesia.
(https://www.muturzikin.com/cartesasiesudest/14.htm)

Expanded Graded Intergenerational Disruption Scale (adapted from Fishman 1991)*			
LEVEL	**LABEL**	**DESCRIPTION**	**UNESCO**
0	International	The language is used internationally for a broad range of functions.	Safe
1	National	The language is used in education, work, mass media, government at the nationwide level.	Safe
2	Regional	The language is used for local and regional mass media and governmental services.	Safe
3	Trade	The language is used for local and regional work by both insiders and outsiders.	Safe
4	Educational	Literacy in the language is being transmitted through a system of public education.	Safe
5	Written	The language is used orally by all generations and is effectively used in written form in parts of the community.	Safe
6a	Vigorous	The language is used orally by all generations and is being learned by children as their first language.	Safe
6b	Threatened	The language is used orally by all generations but only some of the child-bearing generation are transmitting it to their children.	Vulnerable
7	Shifting	The child-bearing generation knows the language well enough to use it among themselves but none are transmitting it to their children	Definitely Endangered
8a	Moribund	The only remaining active speakers of the language are members of the grandparent generation.	Severely Endangered
8b	Nearly Extinct	The only remaining speakers of the language are members of the grandparent generation or older who have little opportunity to use the language.	Critically Endangered
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community. No one has more than symbolic proficiency.	Extinct
10	Extinct	No one retains a sense of ethnic identity associated with the language, even for symbolic purposes.	Extinct

Figure 2. EGIDS language endangerment scale.
(Lewis & Simons 2010)

Background: Languages in Indonesia

Language overview:

- ~725 languages spoken in Indonesia
- Speakers gradually shifting away from local languages
- ~75% classified as *threatened* or lower
- ~10% *moribund* or lower
- Language documentation creates a lasting record of language use
- Language maintenance and revitalization uses documentation to prevent language loss

Background: Linguistic work in Indonesia

Badan pengembangan dan pembinaan bahasa:

- Established in 1948
- Has continually carried out broad linguistic work
- Overseen regional language surveys
- Established first dictionaries and grammars for many languages

Independent research:

- Both community-internal and -external projects
- Led to fine-tuned language output
- Development of language corpora and teaching materials
- Similarly, many first dictionaries and grammars

Background: Linguistic output in Indonesia

Historical resources:

- Earliest sources date back to 17th century
- Large body of wordlists, dictionaries, texts, narratives, ...
- Many language communities unaware of older sources
- Often unavailable because of access, location, or language

Modern language work overlooks many older resources

Electronic access almost ubiquitous, but language work remains behind

Goal: Build upon previous work and prepare current work for future reuse

DATAR: Digitization

Digitization of historical materials:

- Digitally-stored high-quality photos of visual source materials
- Reprocessing of older audio materials into new audio
- Reenters old materials into working databases

Born-digital language projects:

- Resources need not be converted later on
- Access can be immediately made available
- Allows for remote collaboration
- Easier for archiving and version-tracking
- Can always be easily exported to physical materials

DATAR: Annotation

Metadata:

- Rich description of content and its origin
- Makes resources easier to identify and locate
- Provides background information on resource development
- Gives proper attribution to creators and participants
- Useful for tracking and understanding language change

Tools like **Lameta** help manage linguistic metadata

lameta

The Metadata Editor for Transparent Archiving

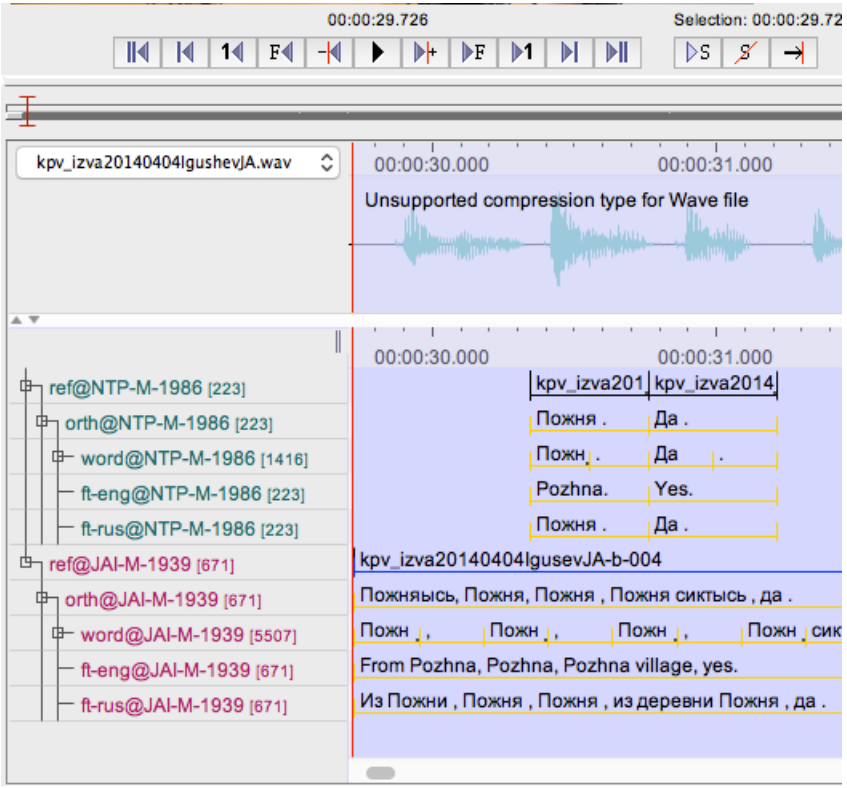

Figure 3. Sample ELAN transcription.
(https://langdoc.github.io/2016-06-04-ELAN-to-R.html)

DATAR: Transcription

Of text:

- OCR or transcription of historical texts
- Materials can be integrated into ongoing projects
- Form the basis of a language corpus

Of audio-video recordings:

- Can be listened back to for language-learning
- Additional translations allow access by wider audience
- Form the basis of a language corpus

DATAR: Archiving

Establishes long-term record of language

Immediately makes resources available to language communities

Allows free and open access to resources

Research and output become replicable

Recent work has seen increase in archival practice

Archives: PARADISEC, TLA, ELAR, Kaipuleohone, **...**

DATAR: Repatriation

Historical materials:

- The export of resources remains a legacy of colonial history
- Many materials remain outside their region of origin
- Repatriation returns rightful ownership and access
- Materials can be reused for future language work

Methods of repatriation:

- Physical return of materials or distribution of copies
- Digital repatriation returns access to creators
- Translation returns them to accessible languages

Future linguistic work

How does it look to use such methods for future linguistic projects?

Digitization: Born-digital project **...**

Annotation: **...** with rich metadata, **...**

Transcription: **...** open-text transcriptions of materials, **...**

Archiving: **...** archived in a public repository, **...**

Repatriation: **...** and available to – and in collaboration with – language communities.

Conclusion

Recent linguistic work has had to start from scratch

- Historically, resources have not been prepared for future work
- Much has been duplicated in terms of documentation
- → Output slower, collaboration more difficult
- → Takes more time and money

With this in mind, we can set the stage now for future language work

Soon, we may no longer be able be able to start from the beginning

Thank you!

Thank **you all** for attending

Questions?

References

Lewis, Paul M. & Gary F. Simons. 2010. Assessing endangerment: Expanding Fishman's GIDS. *Revue roumaine de linguistique* 55(2), 103–120.