

The state of Sumatran lexicography:

Ongoing projects and prospects for lexical documentation

Blaine Billings (blainebillings.com)

University of Hawai'i at Mānoa

DSNA 25, Buffalo, NY, USA

31 May 2025



Roadmap

1. Background
2. Sumatran lexicography
3. Ongoing lexicographic work
4. Broader implications
5. Future lexicographic prospects

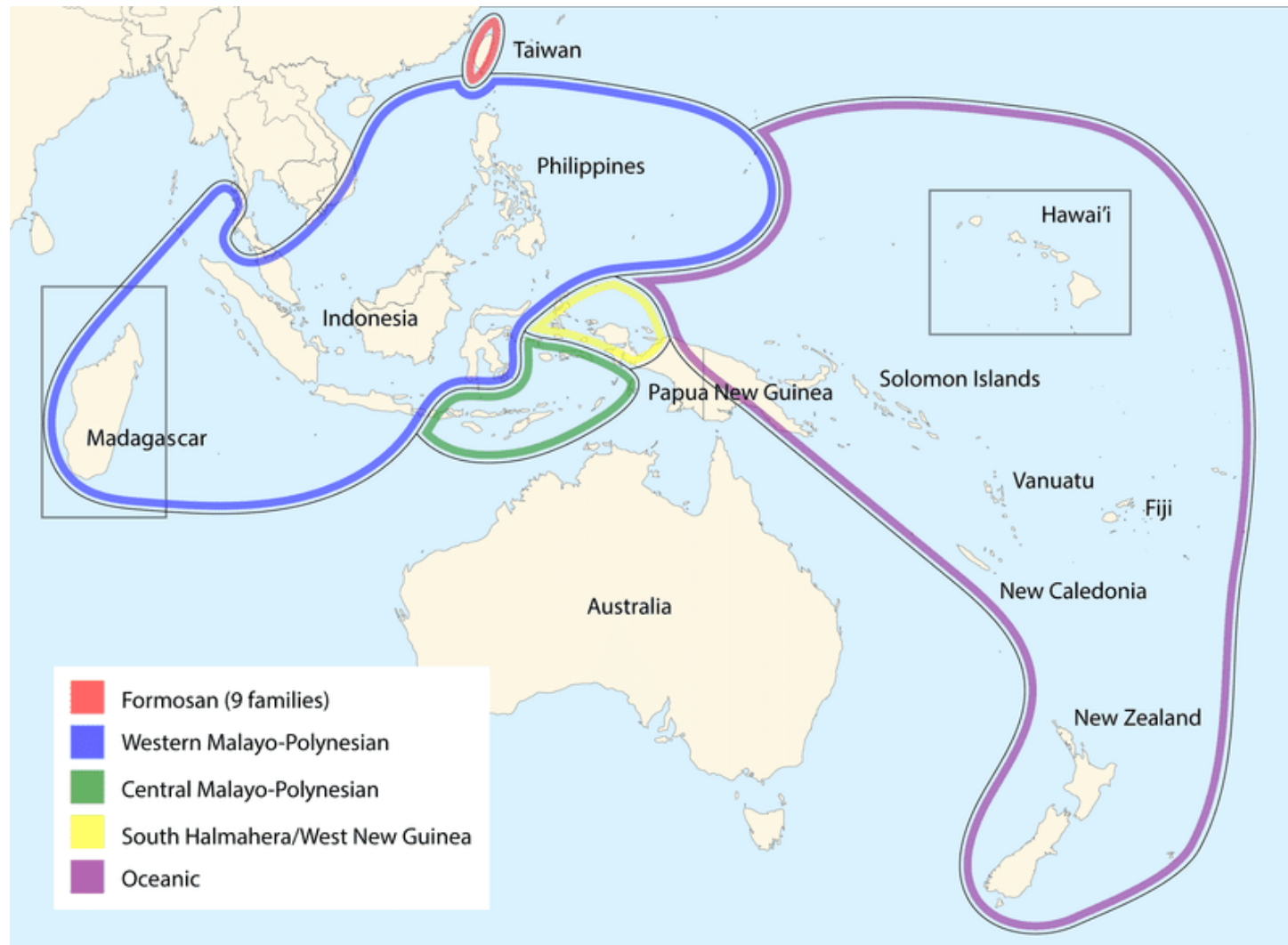


Figure 1. Map of Austronesian subgroups.
(<https://historyguild.org/what-is-the-austronesian-expansion/>)

Background

Sumatran language subgroup

- × Island
- < Malayo-Polynesian < Austronesian
- ~20 languages (7 subgroups)
- Speakers:
 - Total: ~5,000,000
 - ~1,500,000 (Toba Batak)
 - ~3,000 (Nasal)
- Wide variety in vitality, extant documentation, and institutional support

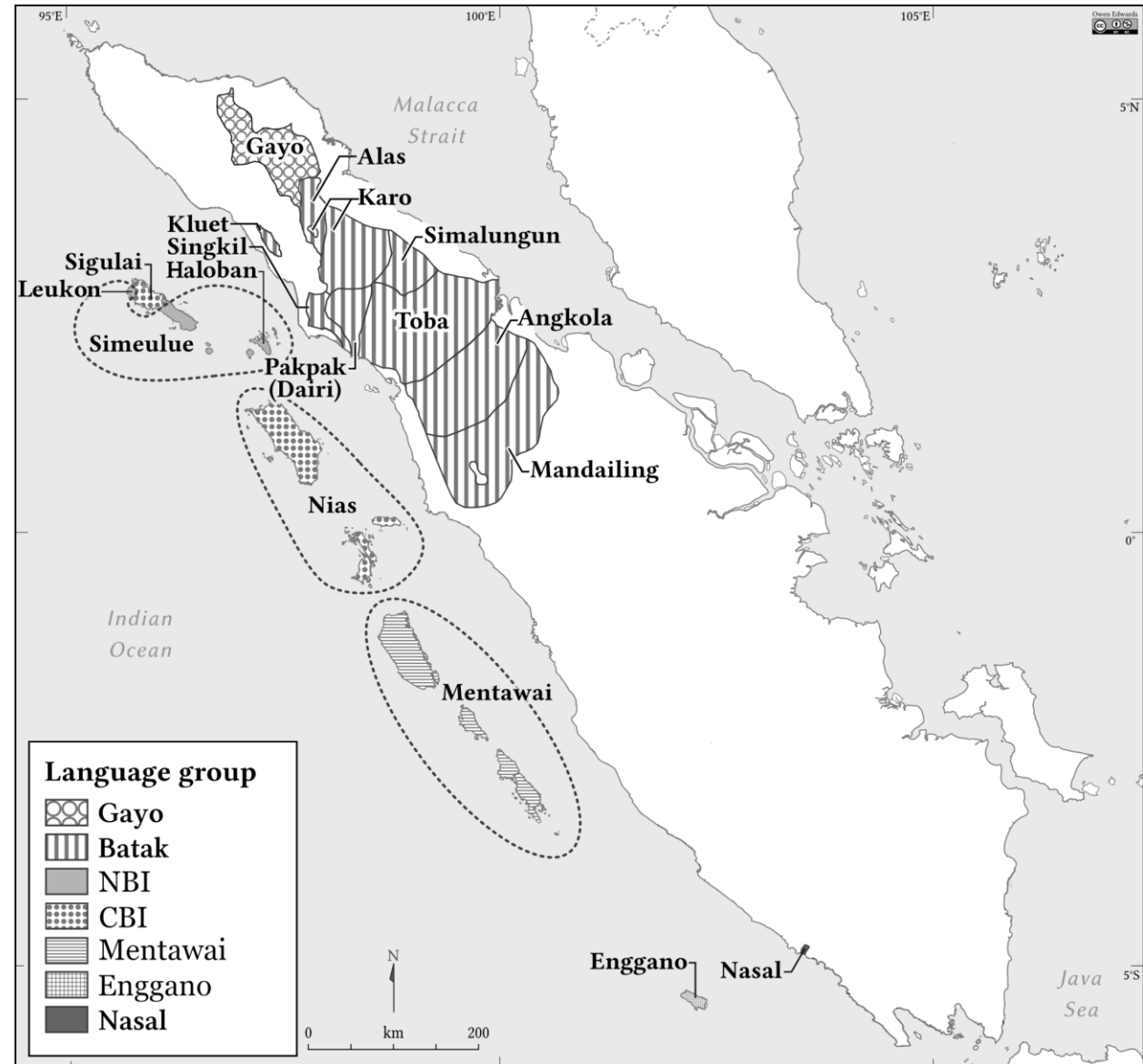


Figure 2. Map of Sumatran languages.

Sumatran lexicography: Precolonial period

Timeframe: <1824

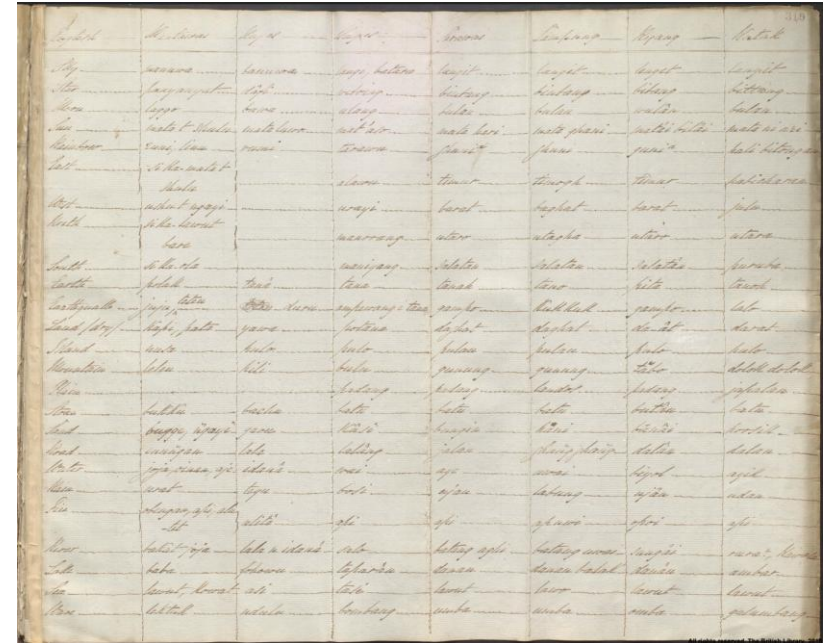
Few scattered documents prior to Dutch colonization

Written material limited primarily to a few wordlists:

- Mostly scattered references in travel notes
- Extensive corpus of Batak script (>2,000 manuscripts)
- No extant written tradition for other languages

Examples:

- Crisp (1799) for Mentawai
- Raffles collection (~1800-1825) for many languages



The image shows a page from a handwritten comparative wordlist. It consists of approximately 10 columns of text, each representing a different language or dialect. The handwriting is in cursive and somewhat faded. The words are arranged in rows, allowing for comparison of related terms across the different languages. Some words are written in red ink, possibly indicating specific grammatical features or roots. The overall layout is organized and systematic, typical of a lexicographical work.

Figure 3. Comparative wordlist from the Raffles collection.

Sumatran lexicography:

Dutch colonial period

Timeframe: 1824–1949

Archipelago-wide wordlist inventory (Holle lists)

Highly detailed descriptions of language use

Language and dialect cartography

Lexicographic resources primarily in Dutch

- **Landmark:** van der Tuuk (1861) for Toba Batak
- **Landmark:** Hazeu (1907) for Gayo

Missionary dictionaries

Other scattered sources

- Travel logs, correspondences, etc.

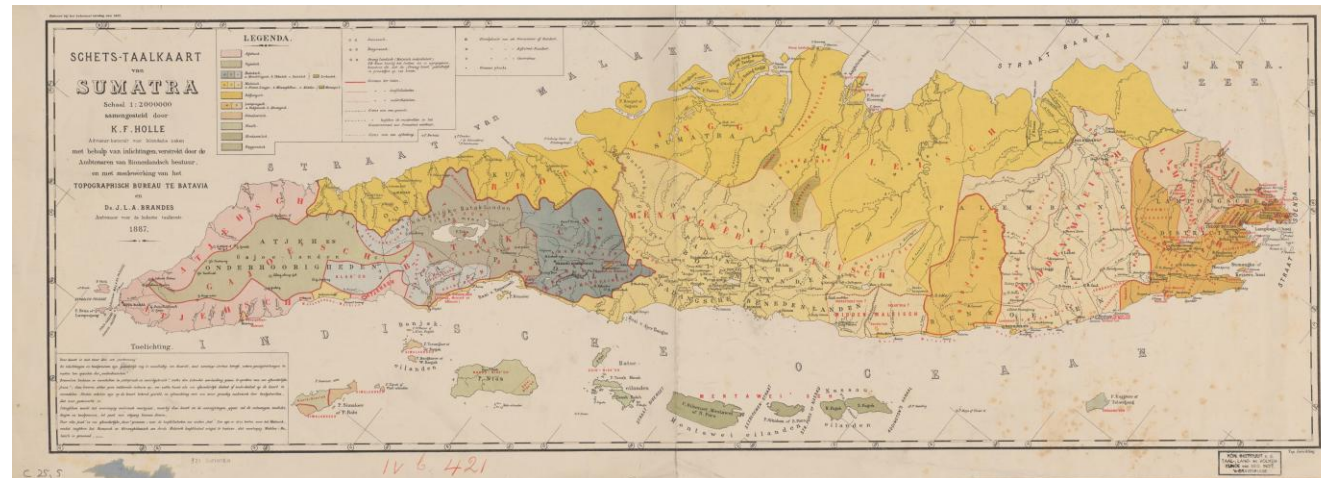


Figure 4. Map of the languages of Sumatra. (Holle 1887)

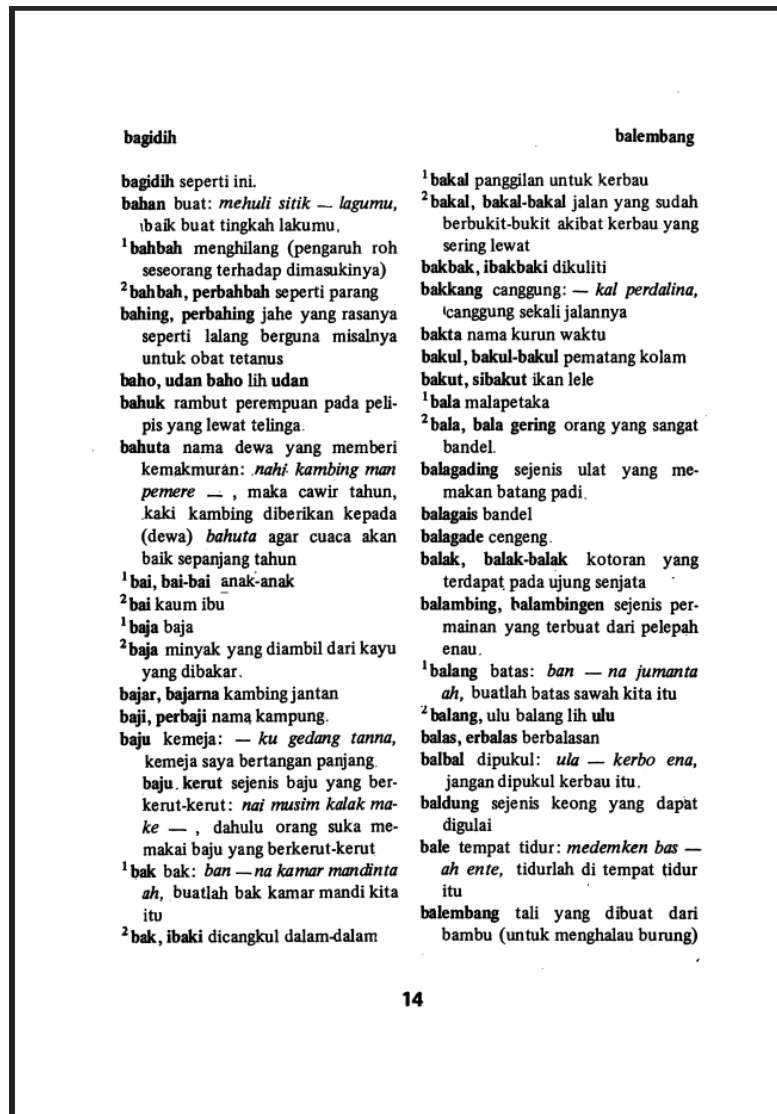


Figure 5. Page from a *PB* Karo dictionary.
 (Pusat Bahasa 2001)

Sumatran lexicography : Indonesian language office period

Timeframe: 1949–present

Indonesia's *Pusat Bahasa* (PB) [Language Office]

- Establishment of regional language offices
- Large-scale description and documentation
- Developed many languages' first dictionaries
- Relatively monolithic
- Bilingual glossing dictionaries

Examples:

- Tim Kamus Balai Bahasa Banda Aceh (2013) for Haloban
- Balai Bahasa Provinsi Sumatera Utara (2015) for Simalungun

Sumatran lexicography:

Current state

Table 1. “Dictionary” resources for Sumatran languages.

Language	# of Senses	Language	# of Senses
Gayo	~24,000 / ~10,000	Mandailing	×
Kluet	×	Leukon	×
Alas	~7,000	Simeulue	~7,000
Singkil	~8,000	Haloban	~4,000
Dairi	~18,000 (~9,000)	Sigulai	~3,000
Karo	~19,000 (~8,000) / ~9,000	Nias	~10,000 / ~10,000
Simalungun	~11,000 (~6,000)	Mentawai	~3,000 / ~3,000
Toba	~25,000 / ~25,000	Enggano	~7,000 (~3,000) / ~4,000
Angkola	~18,000 / ~7,000	Nasal	×

■ = Dutch/German

■ = Indonesian (PB)

■ = Indonesian (Independent)

(#) = Word Roots

Sumatran lexicography:

Current state

Some statistics:

- Languages with a “dictionary”...: 14/18 ⇒ (7 of these <10,000 senses)
- ...in an accessible language...: 12/18
- ...containing primarily definitions rather than glosses: 2/18

Note: I am not aware of *any* monolingual Sumatran-language dictionaries!

“The chances are very slender that this generation or the next will produce more great monuments for any of these regional[=Indonesia] languages.”

— Wolff (1991: 2567)

Ongoing lexicographic work:

Nasal dictionary

Nasal:

- Spoken by ~3,000 people
- Little extant documentary material
- No prior dictionary or written corpus

Collaborative development with community members

Development of first Nasal language dictionary

Primarily from conversational data and subsequent elicitation

To date: ~7,000 entries in ~3,000 word families



Figure 7. A dictionary development session in Nasal.

Ongoing lexicographic work: Sumatran comparative dictionary

Goals:

- Etymological information for Sumatran-language dictionaries
- Reconstruction of Proto-Sumatran vocabulary
- Catalog of sound changes by language

Methods:

- Data from all ~20 Sumatran languages
- Broad representation of lexical sources
- Tagging of inherited and loan cognate sets

To date: Contains ~3,000 cognate sets from ~100,000 entries

Figure 8. (Background) Screenshot of Sumatran comparative dictionary's underlying database.

Hasil untuk lipan

kitaha₁ ark. kitahau

Makna

1) (n) *eng* centipede *ind* lipan

Contoh:

- eno* kah ka'bua' kitaha
ind Mari kita coba lagi dengan seekor kaki-seribu
eng Let us try again with a millipede
source Kähler 1955 Retelling
- eno* dabukai kahai' kitaha, ibuhuade pa ean
ind Mereka menangkap seekor kaki-seribu, dan mereka membiark
eng They got a millipede, and they let it bite it (the child).
source Kähler 1955 Retelling

Ongoing lexicographic work: Barrier Islands project

Arka et al. (2023) Barrier Islands project:

“This project will generate significant new intellectual resources on past and present languages [...] Practical outcomes include [...] language corpora, lexical databases (for bilingual dictionaries), and audio-visual resources underpinning cultural revitalisation.”

Recent development of digital Enggano dictionary

Documentation of the lexicon of Mentawai languages

Support of local language maintenance and vitality

Figure 9. Entry from the digital Enggano dictionary.
(Rajeg et al. 2025)

Broader implications:

Documentary material development

Methodology:

- Lexicographic methodology for documentary linguistics underrepresented
(Lachler & Pankratz 2017: 112; Nabirye 2011: 179–181; P. M. Andersen 2020: 9)
- Dictionary development without existing corpus
- Direct interaction and linking with primary data
- Data sovereignty, archiving, and reusability

Serve as lasting record of language's lexicon

Provide foundation for future lexicographic projects

Broader implications:

Language maintenance

Dictionaries often first resource desired by language communities

Frequently seen as starting point for language maintenance/revitalization

Directly usable by language learners and maintainers

Able to be repurposed into instructional materials

- Database into mini-dictionaries (Mosel 2009, 2011)
- Alphabet and counting books
- Word class and syntactic information for grammar instruction

Broader implications:

Sumatran history

“the historical archaeology of Sumatra [...] has been neglected by researchers in comparison with other islands of the Indonesian archipelago” — Tjoa-Bonatz (2020: xvi)

Comparative reconstruction of vocabulary:

- Determination of Urheimat
- History of agriculture and metalworking
- Role of seafaring and navigation

Analysis of loan vocabulary:

- Subgroup-internal: Diffusion across the region
- Subgroup-external: Arrival of Malays in Sumatra
- Family-external: History of contact with Sanskrit, Tamil, Arabic, Dutch, ...

Future lexicographic prospects

Many Sumatran languages remain without thorough (or any) lexicographic material

Electronic lexicography still largely unexplored

Further study of evolution of lexicographic practice

Global Batak manuscript corpus yet to be cataloged and analyzed

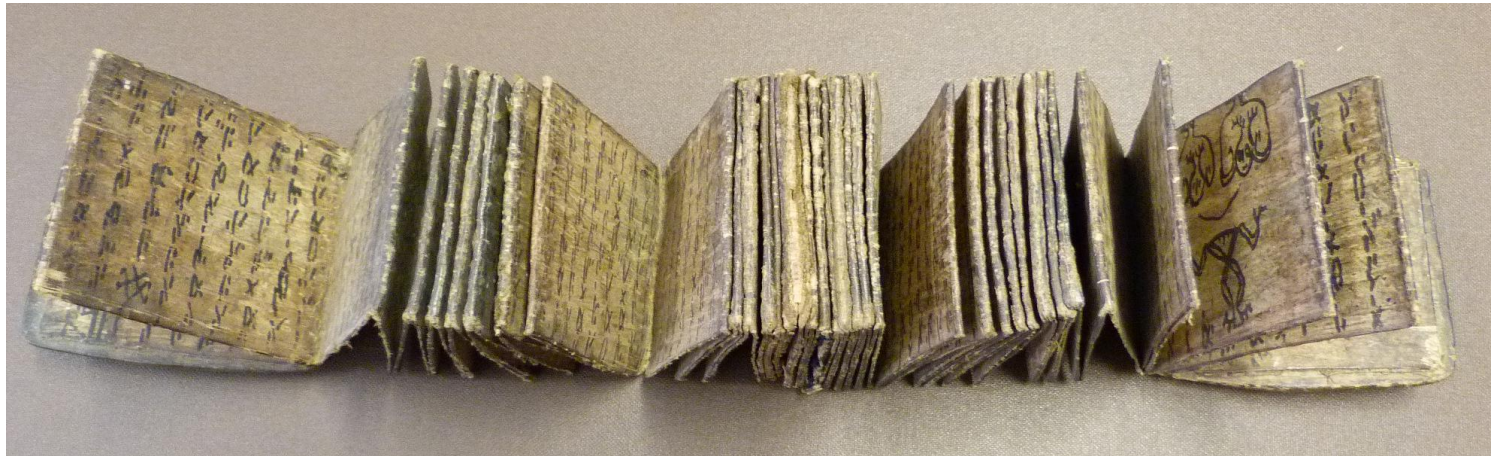


Figure 10. A Simalungun Batak *pustaha* (bark book).

(<https://blogs.bl.uk/asian-and-african/2016/11/batak-manuscripts-in-the-british-library.html>)

References

- Andersen, Patricia M. 2020. *Revitalization lexicography: The making of the new Tunica dictionary*. Tucson: University of Arizona Press.
- Arka, I Wayan, Simon Greenhill, & Dani Or. 2023. *Languages of the Barrier Islands, Sumatra: Description, history and typology*. Project description. (<http://researchportalplus.anu.edu.au/en/projects/languages-of-barrier-islands-sumatra-description-history-and-typo>)
- Balai Bahasa Provinsi Sumatera Utara. 2015. *Kamus bahasa Simalungun–Indonesia*, 2nd edn. Medan: Balai Bahasa Provinsi Sumatera Utara.
- Crisp, John. 1799. An account of the inhabitants of the Poggy, or, Nassau Islands, lying off Sumatra. *Asiatick Researches* 6, 77–92.
- Hazeu, G. A. J. 1907. *Gajōsch-Nederlandsch woordenboek*. Batavia: Landsdrukkerij.
- Holle, K. F. 1887. *Schets-taalkaart van Sumatra*. Batavia: Topographische Inrichting.
- Lachler, Jordan & Elizabeth Pankratz. 2017. Moving towards value-added digital repatriation in lexicography for indigenous languages in Canada. In Nicholas Ostler, Vera Ferreira, & Chris Moseley (eds.), *Proceedings of the 21st FEL Conference*, 107–114. Hungerford: Foundation for Endangered Languages.
- Manik, Tindi Radja. 2002. *Kamus Pakpak Indonesia*. Medan: Bina Media.
- Mosel, Ulrike. 2009. Turning a linguist's lexical data base into a community dictionary. Presented at *1st International Conference on Language Documentation and Conservation*, Honolulu, HI, USA, 12–14 March.
- Mosel, Ulrike, 2011. Lexicography in endangered language communities. In Peter K. Austin & Julia Sallabank, *The Cambridge handbook of endangered languages*, 337–353. Cambridge: Cambridge University Press.

References

- Nabirye, Minah. 2011. Compiling the first monolingual Lusoga dictionary. *Lexikos* 19, 177–196.
- Raffles Collection (1783–1825). British Library: India Office Records and Private Papers.
- Pusat Bahasa. 2001. *Kamus bahasa Karo–Indonesia*. Jakarta: Balai Pustaka.
- Rajeg, Gede Primahadi W., Charlotte Hemmings, Cokorda Pramatha, Engga Zakaria Sanigan, Dendi Wijaya, Sarah Ogilvie, Daniel Krauße, I Wayan Arka, Mary Dalrymple, Bernd Nothofer, Putu Widyantara, Agus Dharma, Tude Maha, Agung Mahadana, & Gung Krisna. 2025. *Kamus digital bahasa Enggano*. Online. (<https://enggano.cirhss.org/>)
- Tjoa-Bonatz, Mai Lin. 2020. *A view from the highlands: Archaeology and settlement history of West Sumatra, Indonesia*. Singapore: ISEAS-Yusof Ishak Institute.
- Tim Kamus Balai Bahasa Banda Aceh. 2013. *Kamus bahasa Haloban*. Banda Aceh: Balai Bahasa Banda Aceh & Yayasan PeNA Banda Aceh.
- van der Tuuk, Herman Neubronner. 1861. *Bataksch-Nederduitsch woordenboek*. Amsterdam: Frederik Muller.
- Wolff, John U. 1991. Lexicography of the languages of Indonesia aside from Indonesian and Javanese. In Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, & Ladislav Zgusta (eds.), *Wörterbücher* [Dictionaries], vol. 3, 2566–2568. Berlin: Walter de Gruyter.

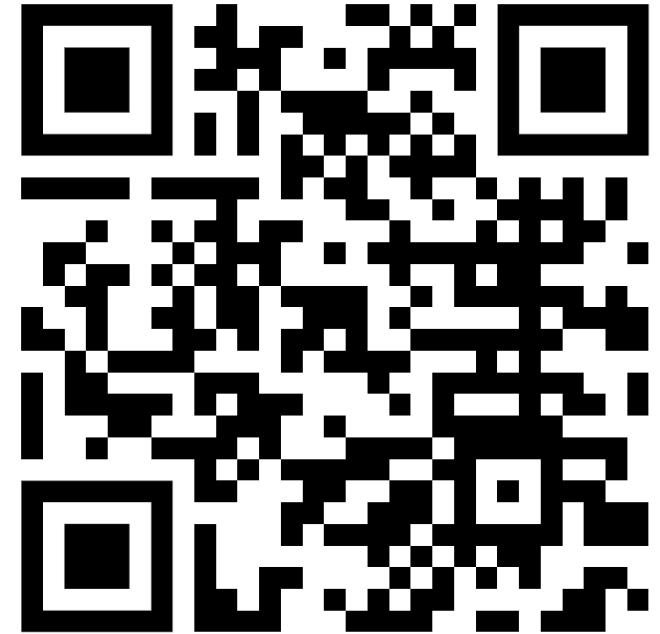
Acknowledgements

My advisor Dr. Bradley McDonnell

Nasal speech community

Thank **you all** for attending

Questions?



↑ Personal Website ↑

(blainebillings.com)