The state of Sumatran lexicography: Ongoing projects and prospects for lexical documentation

Blaine Billings (<u>blainebillings.com</u>) 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

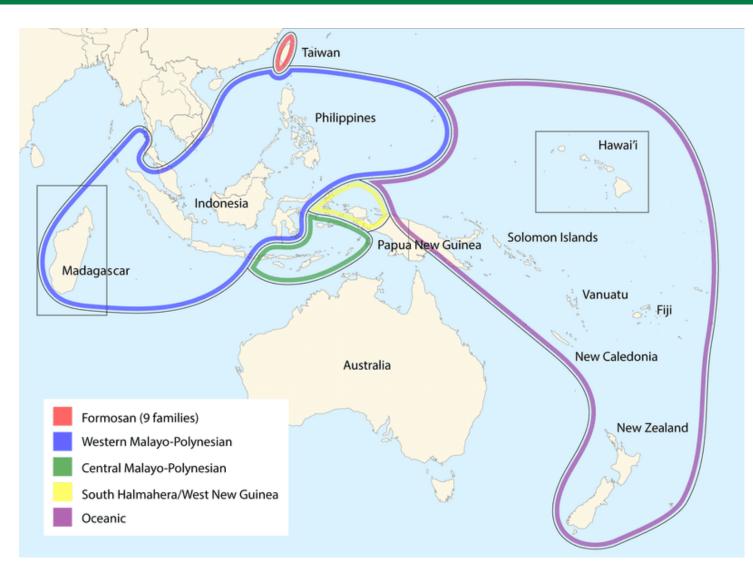


Figure 1. Map of Austronesian subgroups.

(https://historyguild.org/what-is-the-austronesian-expansion/)

Background

Sumatran language subgroup

- \circ × 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

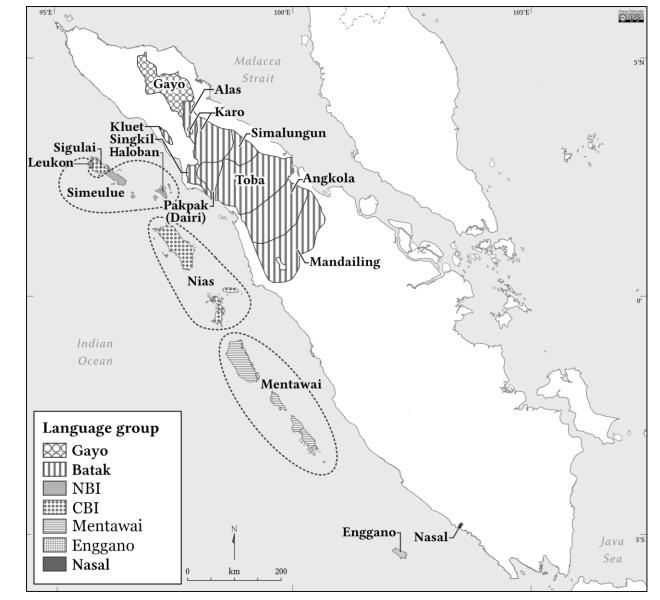


Figure 2. Map of Sumatran languages.

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 <u>many</u> languages

11								Contraction of the local division of the loc
		Menterna	f. at		Anorras	Senfrage_	topang -	Matal
1			lap al		. 200000	-despersion	ropany	Walda
3		Set and	lannard	Gazi Latera	lagit	heapitan	leaget	Lengit
	ther manue	lay any star			bisting	hintang	bilano	bittoing
	les	lappor	tome	alang	Lules	balan	water	halen
	Juin	insta to Mala		est'son	male heri	Inate phani	Inster biller	pato as asi-
	Reinbour	suni lina	manue	Trem	parit	janne	junia	hali bitong a
	Test	tika-water +						
1		Martin 1		alarre	timet	theogh	That	paticharren-
	litertan	whent agazi-		wayi	barat	baghat	heret	Julan
18	hath	tike lawat						
1		bere		hearrang	uter	utesha	ater	atere
1	bett	to harda		taquigang	Jalatan	Inlatan	Selates	hurshe
15	tertte	polad	tand	tere	tanak	tano	hite	taunt
8	Easthemalle	inja taten	the dury		sambour	hall had -	render	lat-
1	Land /dry/-	hafi pate		amperiang = teas		daghal	da at	darat
	Stand		42000	potane	datast			
		halt	helo	hulo	Sular	pulse	Salo	halo
	Monatain	lelen	hili	bula	gunnag	guanag	tabe	
	Alere man	1.00		hadarg	hedrag	lander	hedreg	johalan -
	Atres	butili	bacha	Late	hate	hate	butter	hate
	head	pages upage	1200	hale	haspie	hans	heuse	dordin
	lived	canagan	lake	heling -	jalan	placing placing	dalas	dalan
	Itato	inja vises sje	edana	their man	45	and and	high	spil
	have	unet	tegu	boli	ajace	labung	with	adaa
	his	shaper ali ele	alite	eli -	api		phoi	
		-lit				ghave	2000	y/
Y	herer	falst joja -	lala n idana -	Alt	hating apli -	batang unai-	langer	rune ; Keines
	Jake	heha	fotoara	teparia	tanen	dance balad	danen	ambar
	V42	lacrest, bourst	ali	tali-	levout	laur	lecout	lauration
1	lan	latted	udala	bran hang	unter	unoba	omba	pelantano_

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)

bagidih

balembang

bagidih seperti ini. bahan buat: mehuli sitik --- lagumu. baik buat tingkah lakumu, ¹bahbah menghilang (pengaruh roh seseorang terhadap dimasukinya) ²bahbah, perbahbah seperti parang bahing, perbahing jahe yang rasanya seperti lalang berguna misalnya untuk obat tetanus baho, udan baho lih udan bahuk rambut perempuan pada pelipis yang lewat telinga. bahuta nama dewa yang memberi kemakmuran: nahi kambing man pemere ____, maka cawir tahun, kaki kambing diberikan kepada (dewa) bahuta agar cuaca akan baik sepanjang tahun ¹bai, bai-bai anak-anak ²bai kaum ibu ¹ baja baja ²baia minyak yang diambil dari kayu yang dibakar. bajar, bajarna kambing jantan baji, perbaji nama kampung. baju kemeja: - ku gedang tanna, kemeja sava bertangan panjang. baju kerut sejenis baju yang berkenut-kenut: nai musim kalak make - , dahulu orang suka memakai baju yang berkerut-kerut ¹bak bak: ban - na kamar mandinta ah, buatlah bak kamar mandi kita itu ²bak, ibaki dicangkul dalam-dalam

¹bakal panggilan untuk kerbau ²bakal, bakal-bakal jalan yang sudah berbukit-bukit akibat kerbau yang sering lewat bakbak, ibakbaki dikuliti bakkang canggung: - kal perdalina. canggung sekali jalannya bakta nama kurun waktu bakul, bakul-bakul pematang kolam bakut, sibakut ikan lele ¹bala malapetaka ²bala, bala gering orang yang sangat bandel. balagading sejenis ulat yang memakan batang padi. balagais bandel balagade cengeng balak, balak-balak kotoran yang terdapat pada ujung senjata balambing, balambingen sejenis permainan yang terbuat dari pelepah enau ¹balang batas: ban - na jumanta ah, buatlah batas sawah kita itu ² balang, ulu balang lih ulu balas, erbalas berbalasan balbal dipukul: ula - kerbo ena, jangan dipukul kerbau itu. baldung sejenis keong yang dapat digulai bale tempat tidur: medemken bas --ah ente, tidurlah di tempat tidur itu bälembang tali vang dibuat dari bambu (untuk menghalau burung)

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

¹⁴

baka 2	8 ² balang
 baka baka-baka pelir. bakal jalan berlumpur yg dilalui kerbau dan sebagian juga dibuat olehnya. bakal sej daun yg dipakai utk mencampur candu. bakar habis terbakar: — meseng bernyalanyala. bakat burung gelatik berwarna abu-abu. bakat mendapat luka berat msl oleh pisau atau bambu. bakbak kupas, kuliti; band kolingi. bakit erlalu banyak, sisa; band 'lebih; 	 ²balang membuatnya lari; bakinin tah ~ darohna èna mo asa malot mersora teratoarku èna ataukah km darah itu, maka suling saya ini tidak mau bernyanyi/berbunyi. bakir sej penyakit kepala; bakir-bakiren sakit di sebelah bagian kepala, migren; bakkiren menderita penyakit pd bagian kepala km hujan. bakkang kaku mengenai lidah, tidak dpt berkata dng mudah; band bangkang, jabatan, menetapkan jadi pegawai dsb;
baki-baki sisa-sisa, restan; <i>band</i> iba- iba; sibakina yg lebih, yg ketinggalan, sisa- nya. bakin (<i>khub</i>) oleh sebab, krn: — <i>kuidah</i>	band bangkit, 'naik, tampè. bakko sifat: — kejelman sifat manusiawi; kalak si menguèi — perdagingen me- reka yg hidup dl daging, mereka yg me- nurut kemauan daging.
ngo merpulung-pulung berek i pèa- pèa idi km saya melihat bahwa ikan- ikan kodok berkumpul-kumpul di paya- paya yg dalam itu; — mono gegohmi pakailah tenagamu itu; bakin-bakinen pekerjaan, pekerjaan yg telah dilaksanakan, hal membuat; ibakinken kenca tasak nakan deket ro- roh idi ibakin, ~ mo i perbekasen si Haji Tonggel setelah nasi dan lauk-pauk itu masak, ditaruhnyalah itu di tempat tidur si Haji Tonggel; makinen bakunè ~ gerar pinakan de- ket gerar binatang ? bagai- mana membuat nama ternak dan bina- tang-binatang?; bagainana membuat nama ternak dan binatang-binatang?; memakin membuat, melakukan: keleng atèmu ngo ~ menjadi karinana km ka- sihmu, km kasih sayangmulah, maka semua dpt dilaksanakan, dpt terjadi; pemakinen perbuatan, hal melakukan, perbuatan; pembakin oda ~ tidak dpt dibuat, tidak dpt dilakukan; bakinen, bakinin membuat, melaku- kan;-ku ngo ia merberus saya akan	 Bakon aam amarga. bakong sej bakung: bunga — nama sej kain urk ikat kepala atau destar. bakota mahkota. bakota mahkota. bakta uang jalan, persediaan bahan makanan, an, bekal, perbekalan: enggo keri —ngku dekah aku menepa èna perbekalan saya sudah habis selama menempa ini: band batta, bekel, belanja. ¹baku — siang waktu menjelang fajar. ¹baku ar ³bangku lih bai. bakunè bagaimana, betapa: oda pela — tidak begitu, idad dpt dimaafkan; oda pela — belgahna tidak begitu besar; band makunè katera. bal bersetubuh; band besik, acuk, 'bandel bala sej penyakit yg merusakkan tanamati — pergunggung malim hambis Imam Agung; — pergunggung prajurit ikubu — pergunggung kuta Jerusalen dikepung tentara kota Yerusalem. balak-balak haid; band daroh-daroh. 'balang belalang; band ³baleng. [*]balang halang; band ¹mistak, benè; kimbalang halang band a bandang halang bandang halang bandang halang bandang halang bandang halang hal

Sumatran lexicography : Independent research period

Timeframe: ~1999-present

Community-internal projects:

- Often repatriate/expand on colonial sources
- 10,000+ headwords
- Landmark: Manik (2001) for Dairi Batak

Community-external projects:

- Bilingual/Trilingual w/ Indonesian
- Often result from collaborative documentary linguistic work
- Landmark: Rajeg et al. (2025) for Enggano

Figure 6. Page from Kamus Pakpak Indonesia.

(Manik 2001)

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/Ger	rman = Indonesian (PB)	Indonesian (Independent)	(#) = Word Roots		

Sumatran lexicography: Current state

Some statistics:

- Languages with a "dictionary"...: 14/18
- ...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."

 \Rightarrow (7 of these <10,000 senses)

— Wolff (1991: 2567)

Ongoing lexicographic work: Nasal dictionary

Nasal:

- $\,\circ\,$ 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.

		kicograj		c worl		
Sumatra	n compara	ative diction	ary			
*wil	i (314) *wili (854)			*wili (520) ¶+*aili	*aili¶ - aili	
Goals:						
 Etymologica 	l information for	r Sumatran-language	dictional	ries		
 Reconstructi 	on of Proto-Sum	natran vocabulary				
		*mancil - mai *mancil - ma				
 Catalog of so 	und changes by	language				
Methods:						
• Data from al.	l ~20 Sumatran	languages				
hat • Broad repres	sentation of lexio	cal sources				
 Tagging of in 	herited and loan	n cognate sets				
*did	ak (16: "Dəlak (540" - " - " Ldid (1. *didin (3370					
To date: Cont	ains ~3,000 co	gnate sets from ~ 1	.00,000			
	*bəltaŋ (166	*bəltaŋ ï - mb	,			

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

Indonesia

lipan

 eno dabukai kahai' kitaha, ibuhuade pa ean ind Mereka menangkap seekor kaki-seribu, dan mereka membiarka eng They got a millipede, and they let it bite it (the child). source Kähler 1955 Retelling

Documen Support c

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)

1 May 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 <u>first resource</u> 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:

- $\circ~$ 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

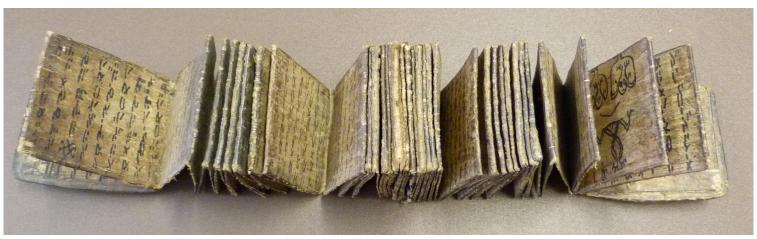


Figure 10. A Simalungun Batak *pustaha* (bark book).

(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. (<u>http://researchportalplus.anu.edu.au/en/projects/languages-of-barrier-islands-sumatra-description-history-</u> 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 Pramartha, 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. (<u>https://enggano.cirhss.org/</u>)
- 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. 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)