Metalexicography for the documentary linguist

Informing internal documentation with external materials

Blaine Billings

University of Hawai‘i at Mānoa

SECOL 2025, Boca Raton, FL, USA

11 April 2025

Outline

Informing

§1 **Metalexicography** for the **Documentary linguist**:

§2 **Internal documentation** with

§3 **External materials**

§4 **Conclusion**

Lexicography

"*The goal of lexicography itself is the creation of dictionaries*"

— Rundell (2012: 63)

"*to present a full account of the words of a language, in all their meanings and patterns of use*" — Kilgariff (2007)

Typical method: Corpus analysis

Typical output: Dictionary

(Meta-)Lexicography

"The development of theories about and **the conceptualization of dictionaries**, specifically with regard to the function, the structure **and the contents of dictionaries**." — Bergenholtz & Gouws (2012: 38, emphasis added)

The goal is a better conception and understanding of lexicographic work

Typical method: Analysis of lexicographic output

Typical output: Subjective remark / theoretical conclusion

Documentary Linguistics

"*an accurate and adequate record of a language for posterity*"
— Rhodes & Campbell (2018: 107)

Documentary linguistics is intended as a wholistic approach

Typical method: Elicitation (+ conversation analysis)
Typical output: Grammar, dictionary, texts

Documentary Lexicography

Documentary Lexicography:

- **Why**: Often the first output desired by language communities
- **What**: Integrates lexicographic methods into documentary context
- **How**: Reconciling goals, outputs, methods, and values

Typical method: Rapid word collection (+ active elicitation)

Typical output:

"about 3,000 entries; by lexicographical standards, this is too short to be considered a full dictionary" — Chelliah & de Reuse (2011: 228)

Documentary Lexicographic Work

3 Typical Stages:

1. Wordlist
2. Expanded wordlist / first-draft dictionary
3. Dictionary project

Each stage of development encounter and address different issues

All three can *accurately* and *adequately* record the language's lexicon

Criticisms of Documentary Lexicographic Work

Primary criticisms:

1. Tend to be *etic* not *emic*
2. Tend to have *blind spots*
3. Overlook semantically *salient domains*
4. *Other*: Errors, accessibility, depth

"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)

Primary Proposal

Given the (lack of) available materials, outside lexicographic work can be leveraged to mitigate these issues

Internal documentation

Internal & External Materials

Internal:

- Extant materials for target language
- Frequently lacking in documentary context
- Usable for language maintenance
- Necessary for linguistic documentation

External:

- Extant materials for other regional/trade languages
- Frequently available at least for language of broader use
- Not mutually intelligible
- Available for future comparative research

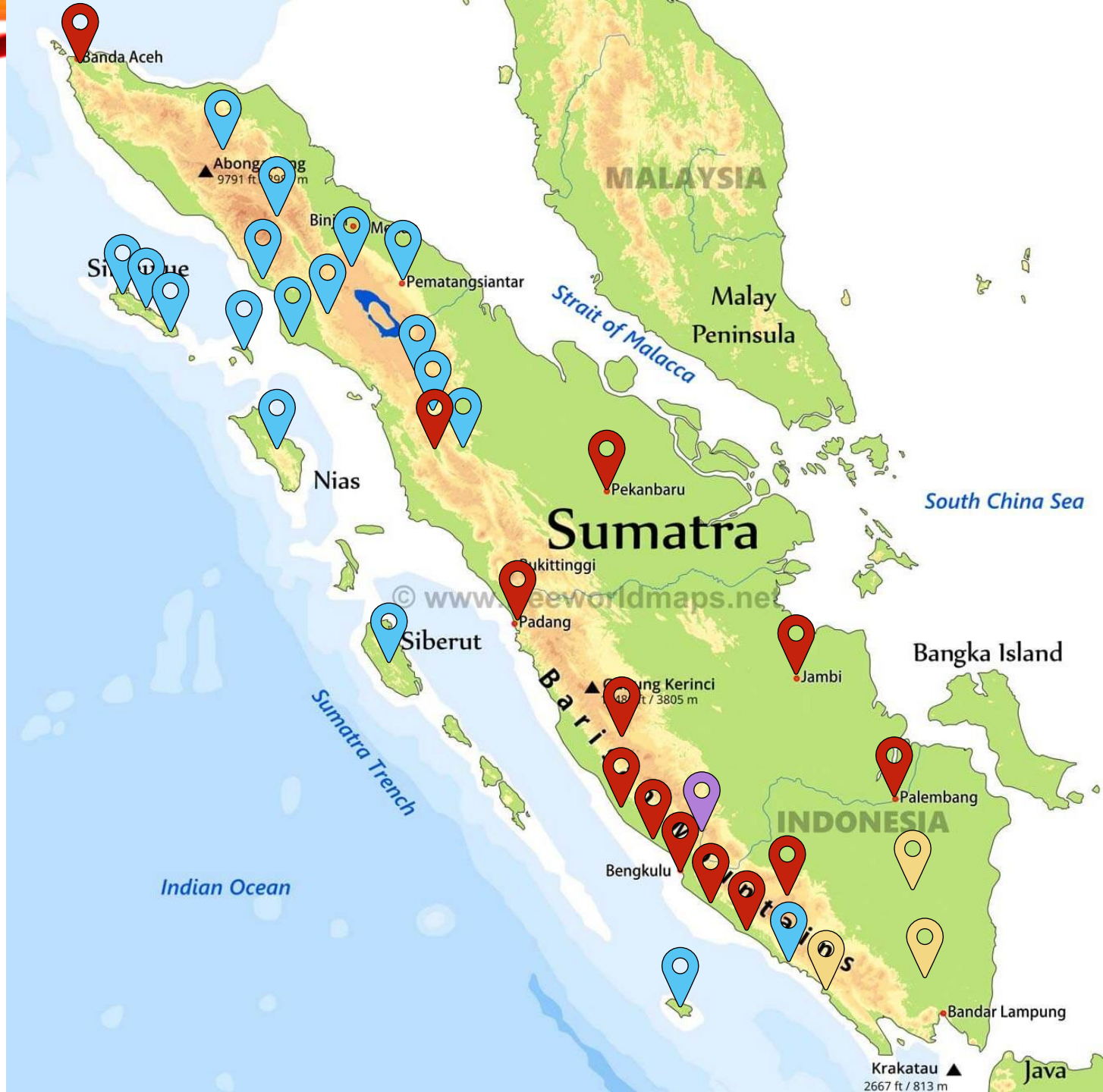
Languages of Sumatra

~35–40+ Languages

Four subgroups of Malayo-Polynesian (Austronesian)

Great diversity in all linguistic domains

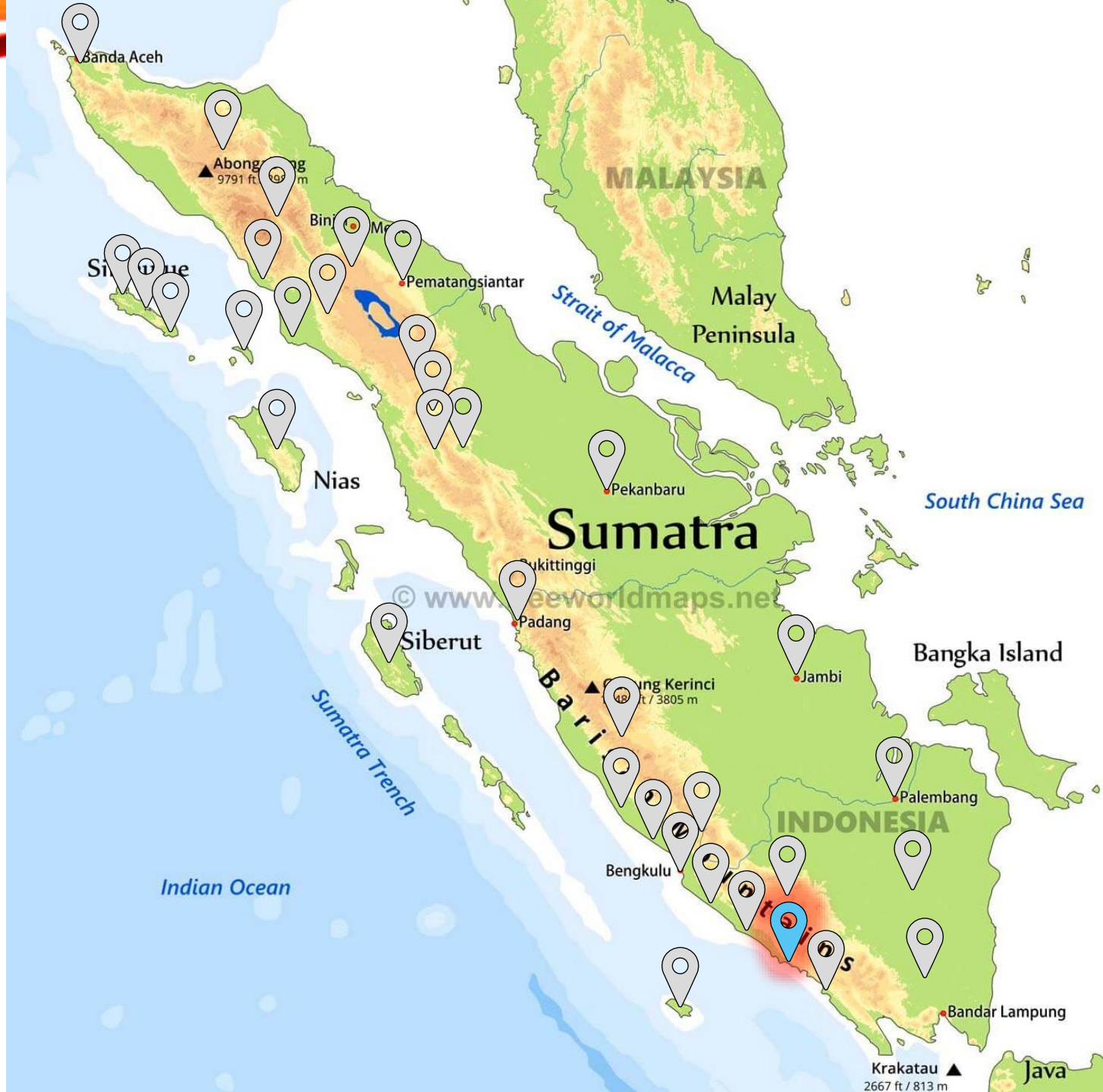
Widely varying levels of documentation



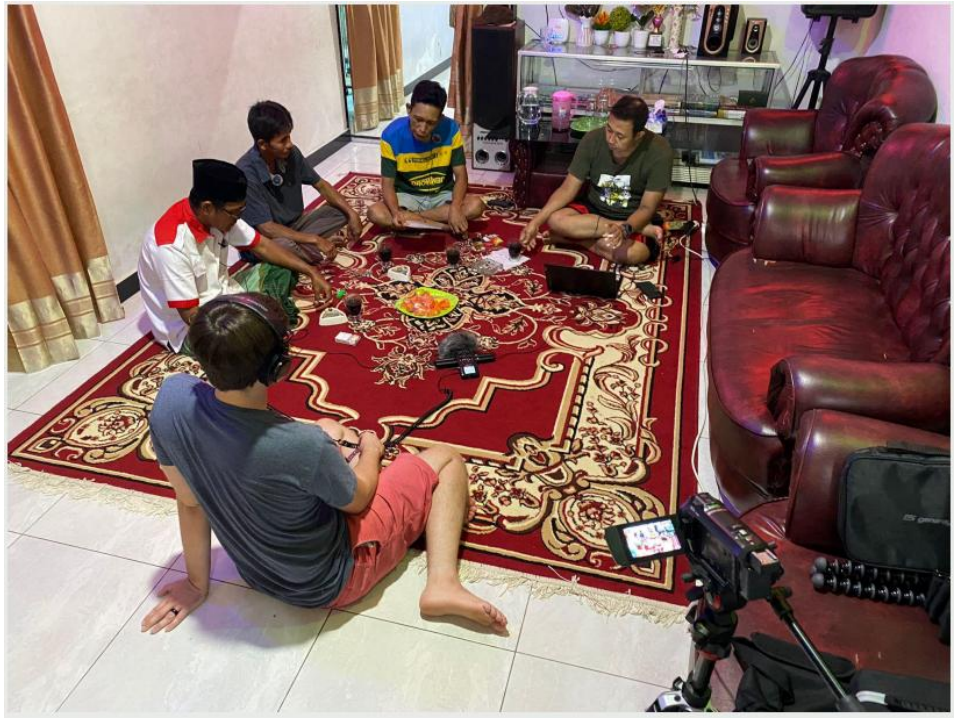
Nasal

Nasal: ✓ [naˈsal] ✗ [ˈneɪ̯.zəl]

- Sumatran < Malayo-Polynesian < Austronesian
- ~3,000 speakers in 3 villages along SW Sumatra
- Unknown to linguists until 2007 (Anderbeck & Aprilani 2013)
- No extant linguistic documentation prior to inception of documentation project
- Endangered with decreasing intergenerational transmission



Nasal Documentation Project



Collaborative effort with members of Nasal speech community

Development of corpus of conversational Nasal speech (McDonnell 2017; McDonnell et al. ongoing)

Ongoing development of Nasal dictionary and readers

Future work on grammar, school textbook, narratives, ...

External materials

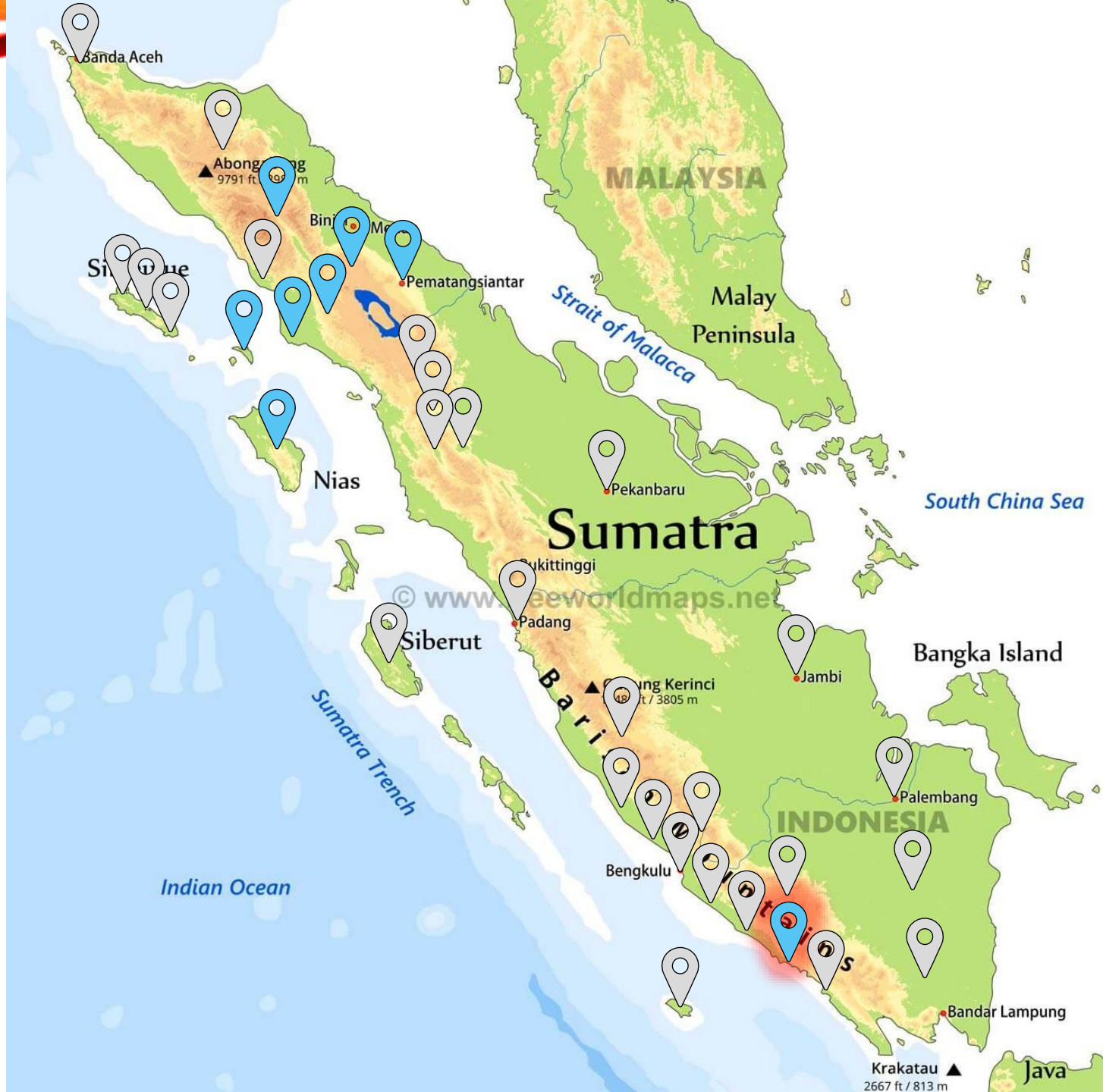
Selection of External Materials

Positives:

- All closely related languages (Sumatran)
- Relative cultural homogeneity
- Convenience

Negatives:

- Overlooks small-scale regional lexical and cultural diffusion
- Slightly differing geographical context



Lexicographic Resources

15 Resources:

- 4 Indonesian dictionaries from Indonesia's language office
- 3 Indonesian dictionaries by independent researchers
- 1 Indonesian mini-dictionary by an independent researcher
- 3 Dutch wordlists
- 2 English wordlists
- 1 Italian wordlist
- 1 English-Indonesian wordlist

3 Sumatran Subgroups:

- Broadest coverage for **Batak** languages, including documentation for Alas, Singkil, Dairi, Karo, Simalungun, Toba
- Dictionary for Haloban (**Northern Barrier Islands**)
- Wordlist for Nias (**Central Barrier Islands**)

Data Compilation

Gathered as part of a separate project

Each source manually transcribed

Each entry tagged with a semantic domain (Moe 2003)

Relative frequencies compiled and compared across tags

~65,000 entries

semdomnum	semdomdesc	page	entry	def-engl-	def-indo-
4.1.9.1.3	Sibling	109	abang	elder brother	abang, kakak laki-laki
8.4.5.1.4	Final	109	abis	last	terakhir
8.4.5.1.4	Final	109	pengabisan	last	terakhir
8.4.5.1.4	Final	109	pengabisen	last	terakhir
8.3.3.3.3	Gray	109	abu-abu	grey	kelabu
8.6.1	In Front	109	adepen	in front of, before	hadapan
7.5.3	To Mix	109	aduk	to stir, mix, meddle	campur, ngaduk
3.2	To Think	109	agak	to think	kira, anggap
3.2.7.1	Hope	109	agak	to hope (?)	harap (?)
3.2	I Think	109	agakku	I think	saya kira
9.6.2.7	So That	109	agar-agar	in order that, so that	supaya, agar
7.4	To Have	109	ajang	to have, possess, own	mempunyai
3.2.2.1	To Study	109	ajar	to study	belajar
3.2.2.1	To Study	109	belajar	to study	belajar
3.2.2.1	To Study	109	telajar	to study	belajar
3.6	To Educate	109	ngajar	to teach	mengajar
3.6	Student	109	pelajar	student	pelajar, murid
8.4.6.1.3	End	109	akér	to finish, end	akhir
9.6.2.6	Result	109	akibat	consequence	akibat
9.2.3	1S	109	aku	I	aku, saya
5.1.1	Mat	109	amak	a mat	tikar
3.4.1.2.2	Calm	109	aman	peaceful, quiet	aman

data_source	data_type	semdomnum	num_tokens	percent_total_tokens
Akbar et al. 1985	dictionary	1.6.1.1.3	18	6.2068965517
AM & Kombih 2012	dictionary	1.6.1.1.3	23	8.1560283688
Manik 2002	dictionary	1.6.1.1.3	39	6.7125645439
Nasal	dictionary	1.6.1.1.3	17	4.4270833333
Saragih et al. 2015	dictionary	1.6.1.1.3	29	7.2139303483
Siregar et al. 2001	dictionary	1.6.1.1.3	33	8.5271317829
Vohry 2016	dictionary	1.6.1.1.3	9	6.976744186
Candrasiah & Kalsari 2018	mini-dictionary	1.6.1.1.3	3	16.6666666667
Soravia 2007	wordlist	1.6.1.1.3	9	13.4328358209
Stokhof & Almanar 1895	wordlist	1.6.1.1.3	11	8.4615384615
Unknown 1822	wordlist	1.6.1.1.3	1	20
von Rosenberg 1855	wordlist	1.6.1.1.3	8	17.3913043478
Westenenk 1904	wordlist	1.6.1.1.3	5	33.3333333333

First Findings: Categories

Semantic:

- Small set of animal classifications
- Large inventory of binomial fruit lexica
- Large inventory of olfactory lexica
- Lunar calendar counting
- Object-dependent eating verbs
- Ocean-oriented movement verbs

Grammatical/Pragmatic:

- Neutral pronoun
- Clusivity in 1P pronouns
- Three-way register (certain vocabulary)
- Three-way deixis
- Multiple negators
- Small set of nominal classifiers
- Specific set of aspectual markers

First Findings: Blindspots

Underrepresentation:

- Mammals, insects, binomial lexica
- *Etic* kinship terms and vocatives
- Deference/taboo vocabulary
- White/black magic
- Traditional religion
- Planting cycle and growth stages

Overrepresentation:

- Large ruminants and birds
- *Emic* kinship terms
- Compositional compounds
- Religion, law, and government

First Findings: Miscellaneous

Compounds:

CHILD [*anak*] + …	Complement
FRUIT [*uwah*] + …	Subordinate
EYE [*mato*] + …	Focal Point
LIVER [*hatai*] + …	Emotion
WATER [*wayil*] + …	Liquid

4-Word Riddles:

Tekhinguk tebatuk, telupo tetinggal.

'Remembered, brought, forgotten, left behind' → A kind of grass

Polysemy:

WATER ~ RIVER	[*wayil*]
TREE ~ PILLAR	[*watang*]
TASTY ~ COMFORTABLE	[*betik*]
HEAD ~ UPSTREAM	× not in Nasal

Deference/Taboo Vocabulary:

Common elsewhere; examples from Nasal:

cuping bumian 'world ear' *or*

behayo 'danger' → Elephant (*gajah*)

beliauan '2S.FORMAL' → Tiger (*ninik*)

Conclusion

Conclusion

Benefits:

- Identify cultural frameworks
- Identify typical blindspots
- Identify salient domains
- Identify regional lexical patterns

— Using external materials can greatly assist our internal documentation —

Acknowledgements

Dr. Bradley McDonnell

Dr. Irwan Abdullah, Ruby Sleemen

Nasal speech community

Thank **you all** for attending!

References

Anderbeck, Karl & Herdian Aprilani. 2013. *The improbable language: Survey report on the Nasal language of Bengkulu, Sumatra*. Dallas: SIL International.

Bergenholtz, Henning & Rufus H. Gouws. 2012. What is lexicography? *Lexikos* 22, 31–42.

Chelliah, Shobhana L. & Willem J. de Reuse. 2011. *Handbook of descriptive linguistic fieldwork*. London: Springer.

Kilgarriff, Adam. 2007. Word sense. In Eneko Agirre & Philip Edmonds (eds.), *Word sense disambiguation: Algorithms and applications* (Text, Speech and Language Technology vol. 33), 29– 46. London: Springer.

McDonnell, Bradley. 2017. *Documentation of Nasal: an overlooked Malayo-Polynesian isolate of southwest Sumatra*. Endangered Languages Archive. http://hdl.handle.net/2196/00-0000-0000-0010-798B-E.

McDonnell, Bradley, Blaine Billings, Jacob Hakim, Johan Safri & Wawan Sahrozi. ongoing. *The languages of the Nasal speech community*. Collection BJM02 at catalog.paradisec.org.au [Open Access]. https://dx.doi.org/10.26278/5f46870d43f29.

Moe, Ronald. 2003. Compiling dictionaries using semantic domains. *Lexikos* 13, 215–223.

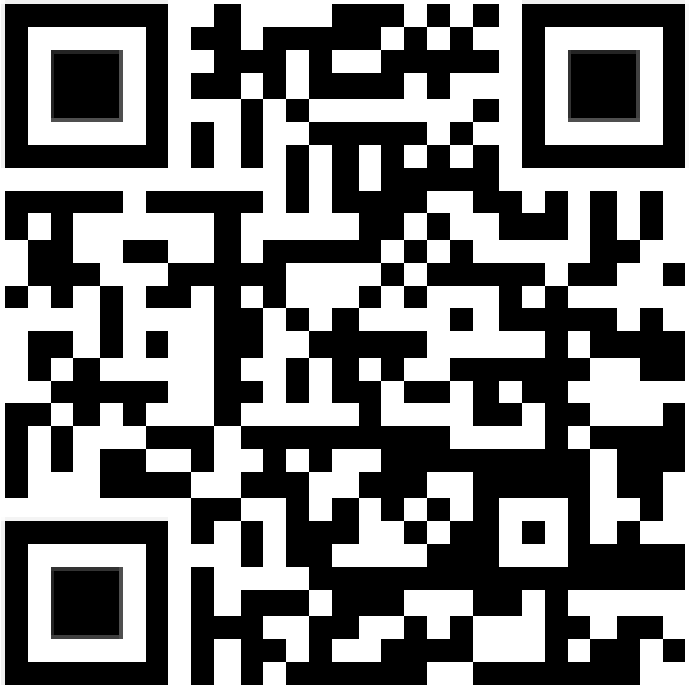
Rhodes, Richard A. & Lyle Campbell. 2018. The goals of language documentation. In Kenneth L. Rehg & Lyle Campbell (eds.), *The Oxford handbook of endangered languages*, 107–122. Oxford: Oxford University Press.

Rundell, Michael. 2012. 'It works in practice but will it work in theory?' The uneasy relationship between lexicography and matters theoretical. In Ruth Vatvedt Fjeld & Julie Matilde Torjusen (eds.), *Proceedings of the 15th EURALEX International Congress*, 47–92. Oslo: Department of Linguistics & Scandinavian Studies, University of Oslo.

Wolff, John U. 1991. Lexicography of the languages of Indonesia aside from Indonesian and Javanese. In Franz Josef Hausmann et al. (eds.), *Wörterbücher* [Dictionaries], vol. 3, 2566–2568. Berlin: Walter de Gruyter

Sumatra Map: https://www.freeworldmaps.net/asia/indonesia/sumatra.html

Questions?



↑ Presentation & website ↑