Building a comparative lexical database for the Sumatran languages

Blaine Billings (blainetb@hawaii.edu)

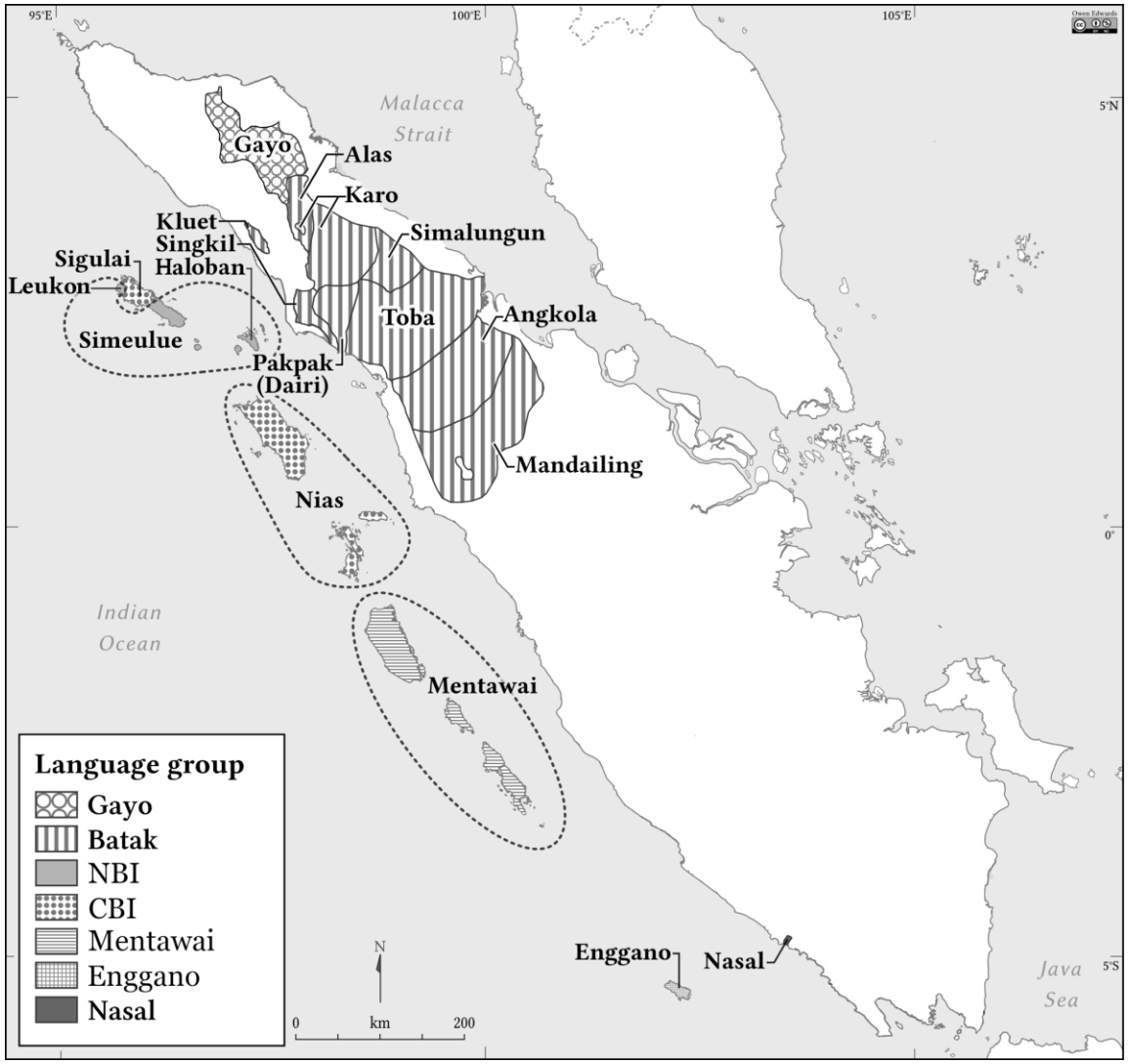
University of Hawai‘i at Mānoa

31st Annual Meeting of the Austronesian Formal Linguistics Association (AFLA 31)

13 June 2024

Outline

1. Background
2. Project
3. Database
4. Finding & Implications



Background: Sumatran

Sumatran languages:

- Subgroup of Malayo-Polynesian (< Austronesian)
- ~20 Languages:
  - **Batak**:
    - **Northern**: Kluet, Alas, Singkil, Dairi, Karo
    - **Central-Southern**: Simalungun,
      - **Southern**: Toba, Angkola, Mandailing
  - **Northern Barrier Islands (NBI)**: Haloban, Leukon, Simeulue (Devayan)
  - **Central Barrier Islands (CBI)**: Sigulai (Sichule), Nias
  - **Other**: Gayo, Mentawai, Enggano, Nasal

Background: Sumatran reconstruction

Notable historical work started with van der Tuuk (1864–1867; 1865)

Nothofer (1986) established “Barrier Island-Batak subgroup”

- Affirmed by Smith (2017)
- Detailed and expanded by Billings & McDonnell (2024)

Adelaar (1981) reconstructed Proto-Batak (PB) phonology and 128 lexical items

Other comparative works:

- Lafeber 1922 (Nias)
- Shorto 1976 (Gayo)
- …

Background: The gap

Little is understood about the early linguistic history of Sumatra

Few lower-level phonological comparisons/reconstructions (Adelaar 1981; Zobel 2021)

No further lexical reconstructions beyond Adelaar (1981) (*that I am aware of*)

~1/2 languages left out of most discussions (data from Blust et al. 2023)

Gayo (189)	**Kluet** (0)	**Alas** (22)	**Singkil** (0)
Simalungun (31)	**Angkola/Mandailing** (10)	**Leukon** (0)	**Haloban** (0)
Sigulai (9)	**Mentawai** (138)	**Enggano** (24)	**Nasal** (0)

Further comparative work necessary for reconstruction (lexical, phonological, …)

Project: Goals

Larger Project: To understand the linguistic and cultural history of Sumatra.

Step 1: Develop a comparative lexicon of the Sumatran languages

- **Refutable**: All reasonings for analytical decisions are cataloged so that they can be argued by other researchers
- **Reproducible**: All information cited can be *easily* identified from their original sources (especially helpful for non-linguistic sources)
- **Reusable**: All compiled and analyzed data is freely available for further research
- **Thorough**: All lexical data is collected and considered, *including* all earlier sources by non-linguists

Project: Methods

1. Transcribe lexical citations (especially the *earliest*)
   - **Lexical**: PMP *sijəm ‘ant’ > PB *sigəm referenced only once in (Kreemer 1922: 209)
   - **Phonological**: Old Mentawai /w/ > Mentawai /b/ in (Oudemans 1879)

Project: Methods

1. Transcribe lexical citations (especially the *earliest*)
2. Identify and catalog cognate sets
   - +Computer-assisted cognate detection

Project: Methods

1. Transcribe lexical citations (especially the *earliest*)
2. Identify and catalog cognate sets
3. Ascribe cognates to inheritance or borrowing
   - *Both* provide insights into linguistic history

Project: Methods

1. Transcribe lexical citations (especially the *earliest*)
2. Identify and catalog cognate sets
3. Ascribe cognates to inheritance or borrowing
4. ⟳ Rinse and repeat with additional sources / languages

Project: Methods

1. Transcribe lexical citations (especially the *earliest*)
2. Identify and catalog cognate sets
3. Ascribe cognates to inheritance or borrowing
4. ⟳ Rinse and repeat with additional sources / languages
5. (*Upload online* & *further research*)

Database: Background

Why a database?

- Portable
- Easily interoperable
- Documentation, support, and functionality

PostgreSQL database

Data entry, editing, and processing in DBeaver

Version control with GitHub (soon to DoltHub)

Database: Structure

Tree-like structure of tables

Three types of tables:

1. *Source* tables:
   - **ID**: Identifies the unique lexical entry
   - **Page**: Page number of source lexical entry is found
   - **Entry**: Orthographic form of entry cited
   - **Definition**: Definition provided for lexical entry
   - **SD Number**: Semantic domain number for entry
   - **SD Description**: 1- or 2-word gloss
   - **Normalized Entry**: Normalized phonological form
   - **Note**: Transcription errors, analytic errors, etc.
   - **...**: Anything else noted by source

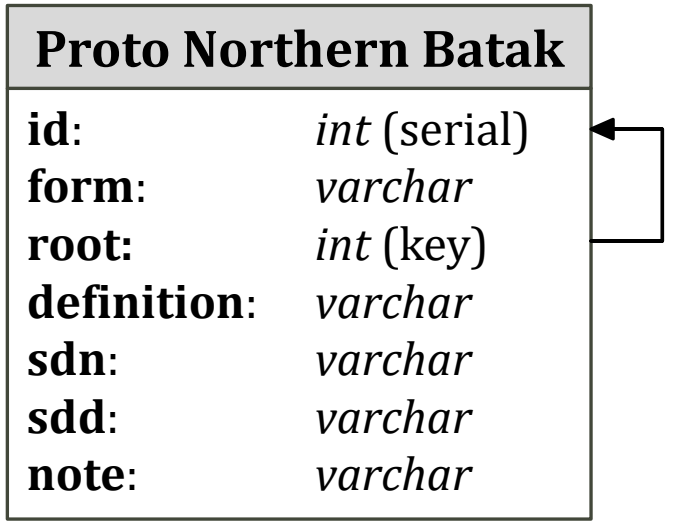
Akbar et al. 1985 (Alas)	
id:	*int* (serial)
page:	*int*
entry:	*varchar*
definition:	*varchar*
sdn:	*varchar*
sdd:	*varchar*
normentry:	*varchar*
note:	*varchar*
...	

Database: Structure

Tree-like structure of tables

Three types of tables:

1. *Source* tables
2. *Reconstruction* tables:
   - **ID**: Identifies the unique reconstruction
   - **Reconstruction Form**: Reconstructed form
   - **Root**: Root of reconstructed form
   - **Definition**: Definition provided for lexical entry
   - **SD Number**: Semantic domain number for entry
   - **SD Description**: 1- or 2-word gloss
   - **Reconstruction Note**: Irregularities, references to other citations, etc.

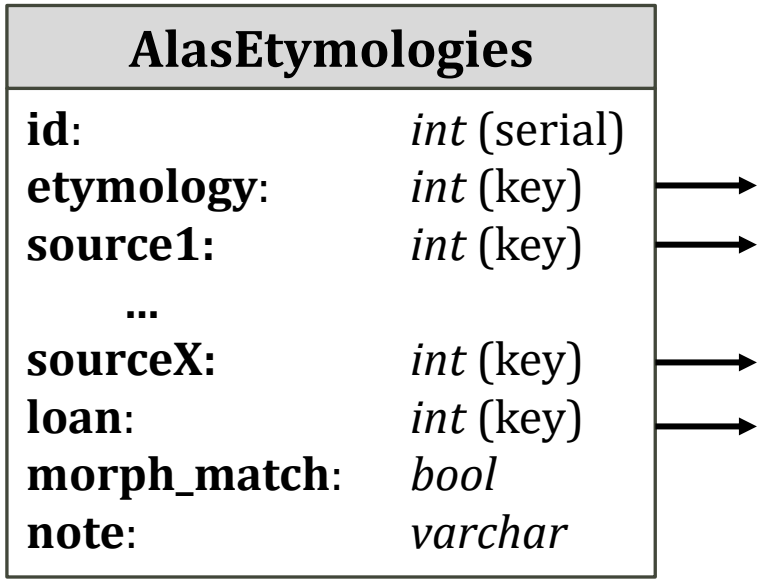


Database: Structure

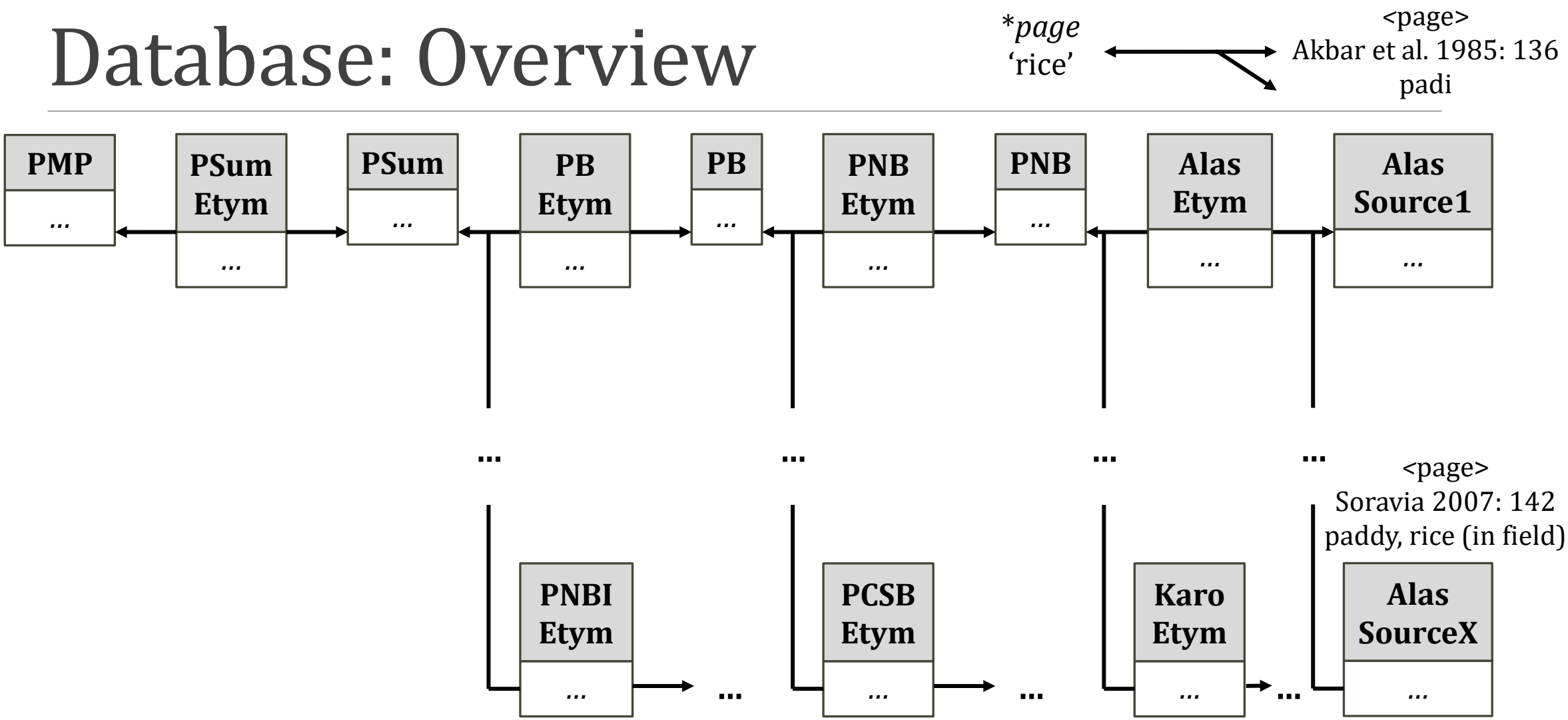
Tree-like structure of tables

Three types of tables:

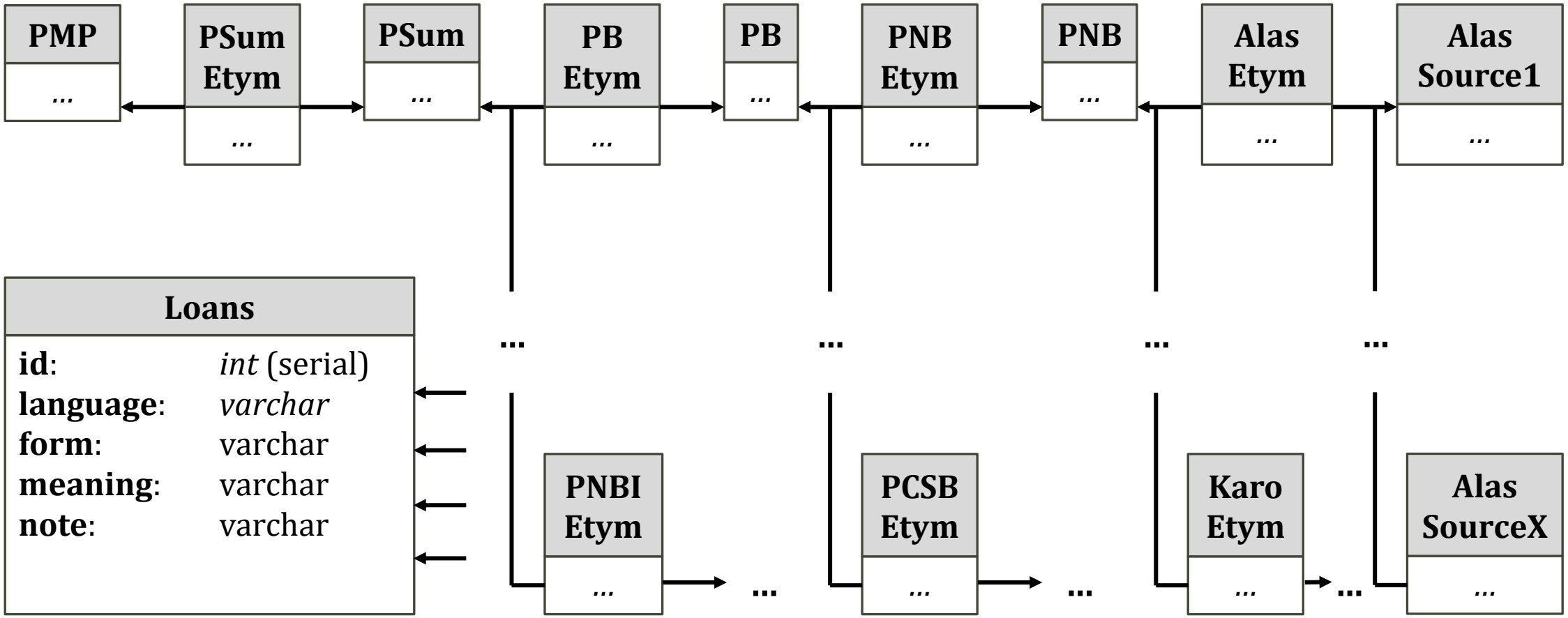
1. *Source* tables
2. *Reconstruction* tables
3. *Etymology* tables:
   - **ID**: Identifies the unique etymology
   - **Etymology**: Reference to reconstructed form
   - **Source 1** … **Source X**: Reference to source forms
   - **Morphological Match**: Whether the morphology of the entry matches that of its etymological source
   - **Etymology Note**: Metathesis, irregular sound change, etc.



Database: Overview



Database: Overview



Database: So far…

Still in *beginning stages*:

- ~65,000 lexical items transcribed from 20 sources (*only* ~30% analyzed so far)
- ~16,000 lexical items ascribed to ~4,000 cognate sets

Most advanced lexical comparison for Northern Batak:

Kluet: 0 → 179	**Alas**: 22 → 1,344	**Singkil**: 0 → 1,422
Dairi: 398 → 1,481	**Karo**: 1,160 → 1,214	**Simalungun**: 30 → 969

Ongoing and next steps:

- → Southern Batak
- → Northern Barrier Islands
- → …

Findings & Implications: Sumatran

Further internal subgrouping:

- Shared Gayo-Batak lexical items: *Closer relationship or just Batak loans?*
- Shared BI lexical items / sound changes: *Closer relationship? Loans? Parallel innovations?*
- Mentawai & Enggano: *Shared lexicon? *s, *t → *t parallel innovation?*

Reconstructed PSum vocabulary:

- Arrival into Sumatra
- Floral and faunal innovations
- Urheimat (maritime or inland)

	Kluet	Alas	Singkil	Dairi	Karo
Liquid metathesis	✓				
*l → y / _#	✓				
*ə → o / _C#	✓				
*ə → e,o / _(l,r)C		✓			
a# mutation	✓ (o)	✓ (ə)			
*ə → o /_h#	✓	✓	✓	✓	
*ə → o / #C_C#	✓	✓	✓	✓	
Monophthongization	✓		✓	✓	✓
*uCə → uCu/oCo			✓	✓	✓
Vowel lowering			✓	✓	
Liquid assimilation			✓	✓	
*lC → rC			✓		

Findings & Implications: PMP

Further evidence for PMP reconstructions (*j, final voiced stops, …)

Evidence for/against PMP phonemes:

- ***c, *z**: Primarily present in loan vocabulary
- ***r, *R**: Distributed *r loans
- ***d, *D**: Distributed *D loans

Possible relationship with other languages of Sumatra: *Rejang? Lampungic?*

Sound changes: *Trisyllable "reduction"? Metathesis?*

Innovations: **wəskir* 'pangolin', **gəmpul* 'bear', **silow* 'otter', (PAN **waNiS* 'boar' > PSum **wili*?)

Findings & Implications: Contact

Intra-Austronesian

- Acehnese loans in Gayo, Batak, NBI: *How far back does contact date?* and *in what domains?*
- Malayic loans in**...** everywhere: *What about maritime vocabulary?* or *Proto-Malayic *əC#?*

Extra-Austronesian

- Sanskrit and Arabic loans: *What domains of use and why?*
  - Sanskrit: Economic, administrative, religious
  - Arabic: Primarily religious
- Additional Tamil/Sanskrit loans not found in Malayic: *More intensive contact?*
- More recent English/Dutch loans: *Why more prevalent in Singkil?*
- Linguistic substrata

Summary

First comparative dictionary for Sumatran languages (*early* work in progress)

Fills a clear gap in Austronesian linguistics

Provides insights into wide variety of aspects of Sumatran (linguistic) history

Beginning of a much larger project in understanding linguistic history of Sumatra

Thank you!

Dr. Bradley McDonnell

Dr. Bob Blust

Dr. Irwan Abdullah, Ruby Sleemen, …

All of you!

References

Adelaar, K. A. 1981. Reconstruction of Proto-Batak phonology. In Blust, Robert A. (ed.), *Historical linguistics in Indonesia*. Jakarta: Atma Jaya University.

Blust, Robert A., Stephen Trussel, & Alexander D. Smith. 2023. *CLDF dataset derived from Blust's "Austronesian comparative dictionary"* (v1.2) [Data set]. Zenodo. doi:10.5281/zenodo.7741197

Kreemer, J. 1922. *Atjèh: Algemeen samenvattend overzicht van land en volk van Atjèh en onderhoorigheden*, vol. 1. Leiden: Brill.

Lafeber, Abraham. 1922. *Vergelijkende klankleer van het Niasisch*. 's-Gravenhage: Hadi Poestaka.

Nothofer, Bernd. 1986. The Barrier Island languages in the Austronesian language family. In Geraghty, Paul, Lois Carrington, & S. A. Wurm (eds.), *FOCAL II: Papers from the fourth International Conference on Austronesian Linguistics* (Pacific Linguistics C-94), 87–109. Canberra: The Australian National University.

Shorto, H. L. 1976. Gayo consonant correspondences. In Đang Liem, Nguyễ n (ed.), *Southeast Asian linguistic studies* (Pacific Linguistics C-42), vol. 2, 119–218. Canberra: The Australian National University. doi:10.15144/PL-C42.199

Smith, Alexander D. 2017. The Western-Malayo-Polynesian problem. *Oceanic Linguistics* 56(2): 435–490. doi:10.1353/ol.2017.0021

van der Tuuk, H. N. 1865. Note on the relation of the Kawi to the Javanese. *Journal of the Royal Asiatic Society* 1: 419–442.

van der Tuuk, H. N. 1971 [1864–1867]. *A grammar of Toba Batak* (Pacific Linguistics Translation Series 13; Jeune Scott-Kemball, trans.). The Hague: Koninklijk Institut voor de Taal-, Land- en Volkenkunde.

Zobel, Erik. 2021. The phonological history of Nias. Paper presented at *Towards the next 40 years of Southeast Asian Studies in Frankfurt: Symposium in honour of Bernd Nothofer*, Frankfurt, 19 December.

Comments/Questions

Extra I: PMP *c

3 Main Sources of /c/:

- PNB Innovations (65/254)
- PB *s → PNB *c / n_ , Ang-Man [nʧ]~[ɲ] (17/254)
  - PB *unsim ‘banana’, *ansəg ‘armpit smell’, *insər ‘kind of fish’, *pansur ‘to spray’, …
- PB *t → PNB *c / _i (27/254)
  - PB *lantiŋ ‘kind of plant’, gətih ‘kind of tree’, pəltik ‘strong’, bərtih ‘roasted rice’, …
- **LOANS**: Malayic, Sanskrit, Tamil (124/254)
- 21 possibly unexplained…?

Extra II: PMP *z

3 Main Sources of /z/:

- **Irregularities**: *qulu ‘upstream’ > *julu*, *quma ‘rice field’ > *juma*
- **Loans**: Malayic, Arabic, Sanskrit, Tamil
- **MORE LOANS**

PMP	*ŋuda	*ŋazəl	*tazəm	×
Malayic	muda	majal	tajam	pijar
Batak	ŋuda	ŋadəl (majəl)	tajəm	pijər

Extra III: Some additional loans...

Tamil பஞ்சி (*pañji*) ‘cotton’ > Batak *panji*

Tamil சூளிக்கு *cuḷikku* ‘spear’ > Batak *culiki* (elsewhere *suligi* < Sanskrit *śūlika*)

Tamil பச்சோந்தான் (*paccōṇāṉ*) ‘chameleon’ > Batak *bindoran* (?)

Sanskrit बृहस्पति (*bṛhaspati*) ‘a god’ > Batak *bəraspati*

Sanskrit छल (*chala*) ‘wickedness’ > Batak cahala

Extra IV: Sample DB entries

PNB

pnb_reconstructions_id	semdomnum	semdomdesc	reconstruction_form
70	5.2.3.1.1	Rice	pagey

Alas Etymologies

alas_etymologies_id	alas_form_etymology	alas_morph_match	alsak1985_reference	alsso2007_reference
1,386	70	[v]	3,349	[NULL]
1,381	70	[v]	[NULL]	914

Akbar et al. 1985

alsak1985_id	semdomnum	semdomdesc	page	entry	def-indo-
3,349	5.2.3.1.1	Rice	136	pagē	padi (tumbuhan yg meng

Soravia 2007

alsso2007_id	semdomnum	semdomdesc	page	entry	def-engl-	def-indo-
914	5.2.3.1.1	Rice	142	pagé	paddy, rice (in field)	padi

Extra V: Example web entry

*page

rice

Language	Entry	Orthography	Definition	Source
PNB	*page		rice	
Kluet	*page*	<page>	padi	(Ismail et al. 1990: 43)
Alas	*page*	<pagé>	padi, paddy, rice (in field)	(Soravia 2007: 142)
		<pagē>	padi (tumbuhan yg menghasilkan beras, juga disebut pagē	(Akbar et al. 1985: 136)
Singkil	*page*	<pagè>	padi	(Vohry 2016: 104)
		<page>	padi, gabah	(AM & Kombih 2012: 73)
Dairi	*page*	<pagè>	padi yg berada di ladang, padi yg belum ditumbuk, jenis-jenisnya	(Manik 2002: 260)