Constructing a multilingual corpus of everyday conversations

Blaine Billings[1], Jacob Hakim[1], Bradley McDonnell[1], Yanti[2], Johan Safri,
Wawan Sahrozi, Hendi Feriza

[1]University of Hawai‘i at Mānoa
[2]Atma Jaya Catholic University of Indonesia

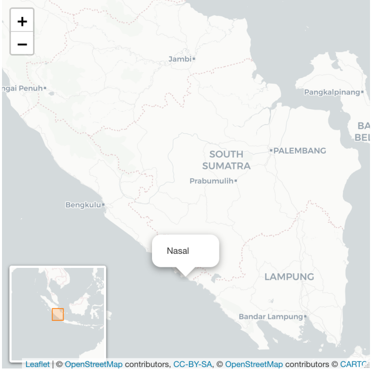
December 16, 2022
14th Association for Linguistic Typology Meeting
Austin, Texas, USA

Overview

Describe on an ongoing project to construct a multilingual corpus of three languages of southwest Sumatra

1. Multilingualism and the Nasal speech community
2. Documenting, managing, and archiving conversational data
3. Case study of variation as an indicator of ongoing change

Nasal and the Nasal speech community



- ≈ 3,000 speakers
  - 3 villages
  - southern Bengkulu province
- Not known to linguists until 2007 (Anderbeck & Aprilani 2013)
- Austronesian language, recently proposed to be a member of the Sumatran subgroup

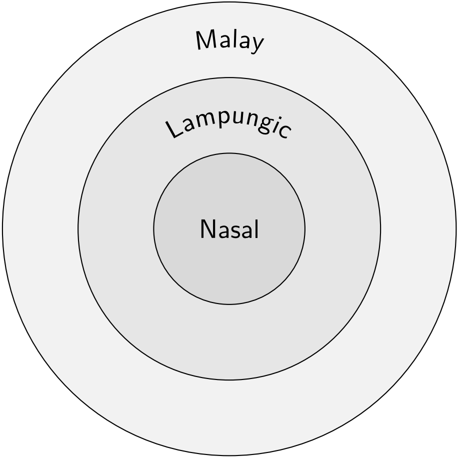
Multilingualism in Nasal speech community

Small-scale multilingualism (e.g. Lüpke 2016)

- Nasal is spoken alongside Malay isolects
  1. Kaur ≈ 40,000
  2. South Barisan Malay (SBM) ≈ 1.5 million (Semende dialect)
- Nasal people speak Kaur and Besemah, but the reverse is often not true
- Nasal shows extreme resilience in light of long-term contact with larger language communities.
- Nasal appears to be at a crossroads where intergenerational transmission is less and less common.

Sustained contact has resulted in apparent restructuring of the grammar

Language Contact and Nasal



Nasal essentially has three layers

1. Nasal (core)
2. Lampungic (earlier influence)
3. Malay (more recent influence)

Complications

1. Malay influence on Lampungic
2. Lampungic influence on Malay (esp. Kaur)
3. Likely not a strict separation Lampungic/Malay influence

Summary of multilingualism

Based on this history of contact and multilingualism, we have set out to document the multilingual speech practices of the Nasal speech community

Documentation of multilingual speech practices

Project goals and desired outcomes

- Documentation, description
  1. Corpus of conversation, culturally important events
  2. Grammar, including detailed description of prosody
- Creation of language resources for language maintenance
  1. Dictionary
  2. Storybooks
  3. Educational materials
- Training and capacity building

Training and capacity building

- Collaborative project
  - Nasal, SBM speech community members
  - Faculty at UNIKA Atma Jaya
  - Faculty/Grad students at UH Mānoa
- Documentation and development of language materials
- Training and capacity building for speech community
- Research into multilingual speech practices

Unfortunately been slowed by the pandemic :(

Audiovisual Documentation: Pre-/Mid-Pandemic

- Pre-Pandemic
  - UH team members travel to Nasal
  - In-person video and audio recording
  - Collaborative work between UH and Indonesian team members
- Mid-Pandemic
  - Team meetings and discussions of research activities via WhatsApp
  - Remote recording via Skype, WhatsApp, Zoom
  - Independent audio/video recording by Nasal team members

Speech Events

- Spontaneous
  - Conversations between community members
  - Narratives
  - Community meetings
  - Cultural events (Sengkukho, Cik Sitian)
- Elicited
  - Discussions of linguistic structure (phonology, morphosyntax)
  - Speaker judgments
  - Phonetic experiments
  - Semantic domain elicitation

Multilingual Speech Practices

Pre-pandemic goals were specifically oriented around collecting Nasal speech data

- Corpus composed mostly of conversations between native Nasal speakers
- Some conversations in Kaur between native Kaur speakers

Moving forward, in the next phase of the project we aim to record more diverse groups of speakers and document the multilingual interactions and dynamics of the Nasal speech community

- Conversations featuring speakers of different languages

Transcription & Annotation

- Segmented into Intonation Units
- Simplified UCSB-style transcription system
- Typical ELAN-FLEx-ELAN workflow
- Additional tiers use to mark discourse transcription, code switching, and conversational context
- Issues:
  - FLEx handling of discourse transcriptions
  - FLEx prioritization of single language
  - IUs vs full utterances
- Currently in first steps of using computer-assisted transcription with Automated Speech Recognition (ASR) software

Data Management

- All materials archived on PARADISEC (Pacific and Regional Archive for Digital Sources In Endangered Cultures)
- FLEx for ongoing dictionary and transcribed conversations
- Lameta for storing metadata
- GitHub for transcription file version control, concurrent file editing and review, and tracking file progress

Corpus

1. Nasal
    1. Recorded: ≈ 55hrs
    2. Transcribed: ≈ 37hrs
2. Kaur
    1. Recorded: ≈ 7.5hrs
    2. Transcribed: ≈ 4.5hrs
3. South Barisan Malay (Semende dialect)
    1. Recorded: ≈ 6hrs
    2. Transcribed: ≈ 5hrs

Already a large corpus of the Besemah dialect of SBM from McDonnell.

Corpus Study

Corpus Study

1. Widespread pattern of variation in the choice of relativizer:
   - *sai* – considered by speakers to be the Nasal form
   - *yang* – Kaur borrowing
2. Similarly widespread variation in 3S pronoun
   - *yo* – considered by speakers to be the Nasal form
   - *nyo* – Kaur borrowing

Corpus Study: *sai* vs *yang*

(1) ***perusahaan sai di pusat**,*
business REL LOC center,

‘A company in the center’,

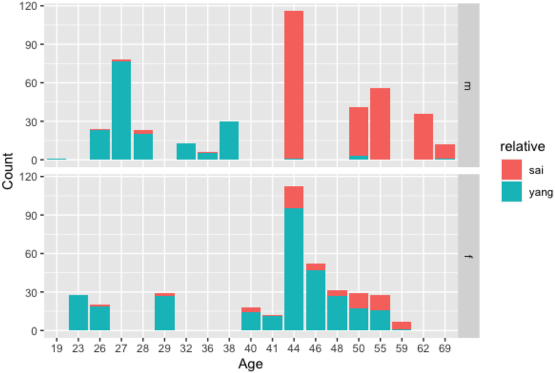
(2) ***yang jelas** kekhejo kuli mudil sudi, ndang te-beban.*
REL clear work laborer type that don’t INV-burden,

‘what’s clear is that (for) that kind of laborer work, don’t be burdened.’

Corpus Study: *sai* vs *yang*

Data analyzed from 9 transcribed conversations

- 14 men and 12 women
- Ages 22 to 72
- Fully code-switched utterances excluded



Corpus Study: *sai* vs *yang*

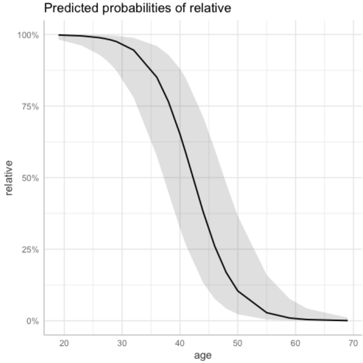
Mixed Effects Model

- Fixed Effects: Age x Gender, Headedness
- Random Effect: Participant

Significant effects:

- Age
- Gender
- Headedness

*Interaction between Age x Gender not significant



Corpus Study: *sai* vs *yang*

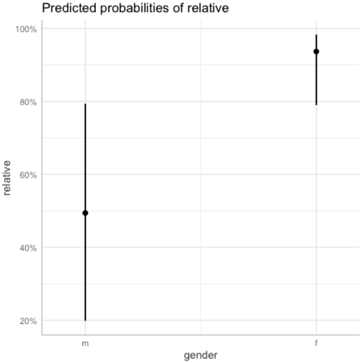
Mixed Effects Model

- Fixed Effects: Age x Gender, Headedness
- Random Effect: Participant

Significant effects:

- Age
- Gender
- Headedness

*Interaction between Age x Gender not significant



Corpus Study: *sai* vs *yang*

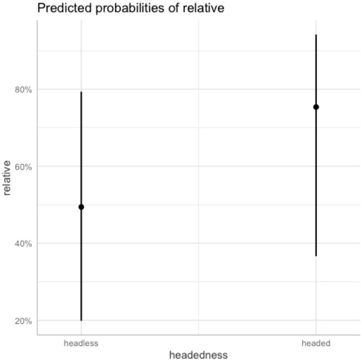
Mixed Effects Model

- Fixed Effects: Age x Gender, Headedness
- Random Effect: Participant

Significant effects:

- Age
- Gender
- Headedness

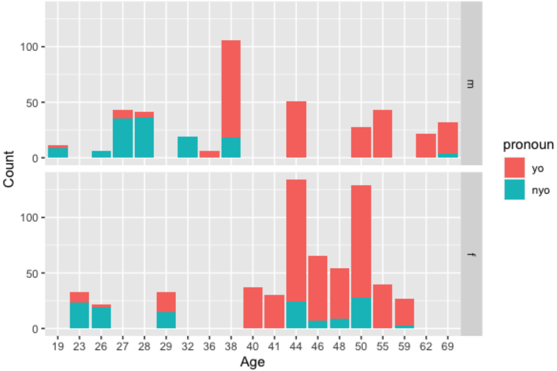
*Interaction between Age x Gender not significant



Corpus Study: *yo* vs *nyo*

Data analyzed from 9 transcribed conversations

- 14 men and 12 women
- Ages 22 to 72



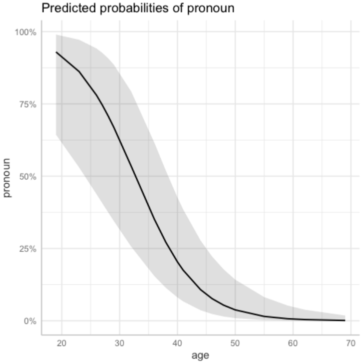
Corpus Study: *yo* vs *nyo*

Mixed Effects Model

- Fixed Effects: Age x Gender
- Random Effect: Participant

Significant effects:

- Age
- NOT Gender



Conclusion

1. Collaborative planning, capacity building and remote teamwork led to the development of a multilingual corpus of conversational recordings of the Nasal speech community
2. Emphasis on collaborative transcription and robust data management
3. Corpus studies using data from multilingual interactions give insight into the mechanisms of language change in progress in Nasal

Acknowledgements

We would like to thank the following organizations and people.

1. The Nasal community from Tanjung Betuah, Gedung Menung, and Tanjung Baru, especially Anton Supriyadi.
2. Kementerian Negara Riset dan Teknologi (RISTEK) for allowing us to conduct research in Indonesia
3. Center for Language and Culture Studies, Atma Jaya Catholic University of Indonesia.
4. Indonesian Institute of Sciences (LIPI), especially Anggy Denok Sukmawati for earlier work on this project

This material is based upon work supported by the National Science Foundation under grant BCS-1911641 to the University of Hawaiʻi at Mānoa. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Anderbeck, Karl R. & Herdian Aprilani. 2013. *The Improbable Language: Survey Report on the Nasal Language of Bengkulu, Sumatra*. Electronic Survey Report 2013-012. SIL International.

Lüpke, Friederike. 2016. Uncovering Small-Scale Multilingualism. *Critical Multilingualism Studies* 4(2). 35–74.