

## Data processing and its impact on linguistic analysis

Anna Margetts & Andrew Margetts  
Saliba/Logea Project

## The project

- Saliba/Logea: Oceanic language spoken in Milne Bay Province of Papua New Guinea.
- The project: web-based database of texts with text-audio linkage and searchable annotations

## Today's talk

- Impact of data treatment on linguistic analysis
- Demonstrate on the basis of two topics:
  - the positioning of PPs
  - change in the use of the plural marker

## Data flow

- Recording
- Capturing to disk & making Premiere file
- Identifying & cutting sessions
- In Transcriber: "Chunking" text into **intonation units** & creating **audio linkage**
- Transcribing text (RA in PNG)
- Converting into **Toolbox/Elan** file
- Translating into English
- (Interlinearizing)

## Dataset used for this talk

IUs	14850
Speakers	62
Texts	87

(All texts that were transcribed at that point)

## Saliba/Logea postpositions

- General postposition *unai* can mark different semantic roles:  
  
location, goal, addressee, recipient, source, instrument

## Location

- (1) *Simai wa ye tu-tuli numa ne unai.*  
cat ANA 3sg RED-sit house DET PP.SG  
'The cat is sitting in the house.'

## Goal

- (2) *Ya lage Samarai unai.*  
1SG arrive Place.Name PP.SG  
'I arrived on Samarai.'

## Order is not fixed

- (3) Figure Verb Ground PP  
*Simai wa ye tu-tuli numa ne unai.*  
cat ANA 3SG RED-sit house DET PP.SG  
'The cat is sitting in the house.'

- (4) Figure Ground PP Verb  
*Simai wa numa ne unai ye tu-tuli*  
cat ANA house DET PP.SG 3SG RED-sit  
'The cat is sitting in the house.'

## Research questions

- Do PPs more commonly precede or follow the verb?
- What is the typical position of PPs expressing different roles?

## Problem: clause boundaries

- When PP is **preceded and followed** by a verb (plus subj pronoun) it is not always clear which clause the PP belongs to and what role it expresses.
- Because:
  - the postposition does not semantically specify the role of its object,
  - clauses can follow each other without indication of coordination or subordination.

## V PP V

*He came up **Magehao unai** he asked*

- (5) *Ye saema Magehao unai ye henamai*  
3SG come.up Place.Name PP.SG 3SG ask  
(a) '[He came up to Magehao] and asked.'  
(b) 'He came up and [at Magehao he asked].'

(WekuSinibu\_01AC\_123-24)

## Text-only database

- Problems: info about pauses, intonation
  - Transcription with pauses may resolve ambiguity
- (6) *Ye saema Magehao unai # ye henamai*  
3SG come.up Place.Name PP.SG 3SG ask  
'[He came up to Magehao] and asked.'

## Text-only database cont.

- Only helpful if pauses preceding OR following the PP.
- Ambiguity remains if no pauses or if pauses before and after PP.
- Reliance on transcription:  
if pauses are missing => wrong analysis

## Text-audio linkage

- Investigate pauses and intonation patterns directly.
- Transcription is a form of analysis, not neutral representation of speech event.

## Unexpected finding

- Even with text-audio linkage, some of the Saliba/Logea examples of PPs remain unclear.
- PPs can be preceded and followed by a verb within the same intonation unit

## Vagueness

- Cases where the PP's association with preceding or the following clause is vague (rather than ambiguous).
- Plays two roles simultaneously:
  - goal of motion verb
  - location of event described by second verb.

## Vague examples

(7) *Se laoma numa wa unai se gwau*  
3PL come house ANA PP.SG 3PL heap  
'They came **to the house** and gathered (things)  
**in the house.**' (Giyahi\_01AA\_073)

(8) *Ye lao unai ye keno-keno*  
3SG go PP.SG 3SG RED-sleep  
'He went **there** and was resting **there.**' (TBLaki\_01AG\_221)

## Relevant data treatment

- Text-audio linkage
- Concordance
- Manipulation of Toolbox to
  - allow jumps from concordance to record
  - play audio of single records in Toolbox

## Plural marking in Saliba/Logea

- Number marking only on nouns with human referents

(9) *wawaya*      *wawaya-o*  
child            child-PL  
'child'          'children'

(10) *wawaya-o*      *gagili-di*  
child-PL            small-3PL.POSS  
'small children'

## Plural marking

- People state that younger speakers tend to use the plural suffix in contexts where it is ungrammatical to the ears of older speakers.
- For example
  - NPs denoting non-human referents
  - On modifiers in NPs with human referents

## Unexpected findings

- Older and younger speakers do it.
- Logea speakers do it less, or in more restricted contexts.
- Artifact of database?

## Data sample

- 60% IUs by Saliba speakers
- 40% IUs by Logea (& Gonubalabala) speakers
- 19 examples of novel plural use in total
- 2 of them by Logea speakers

## Relevant data treatment

- Manipulation of Toolbox
  - to show speaker ID in concordance
  - to jump from concordance to Toolbox record
- Extracting information on speaker's age
- Extracting information on speaker's dialect

## Important points

- Text-audio linkage for direct access to basic data
- Ability to extract info about the database (ideally relational database!)
- Without this treatment
  - analysis would not have been possible OR
  - analysis would have lead to wrong answers