

# Beyond a means to an end:

A case study in building phonotactic corpora for Central Australian languages



Saliha Muradoğlu, James Gray, Jane Simpson, Michael Proctor and Mark Harvey

Image credits: James Gray

# Motivation

- Exploring the phonotactics in four Central Australian languages:
  - Specifically, vowel distributions across syllables
  - Kaytetye, Pitjantjatjara, Warlpiri and Warumungu
    - All Pama-Nyungan languages.
    - Refer to paper for a more detailed overview

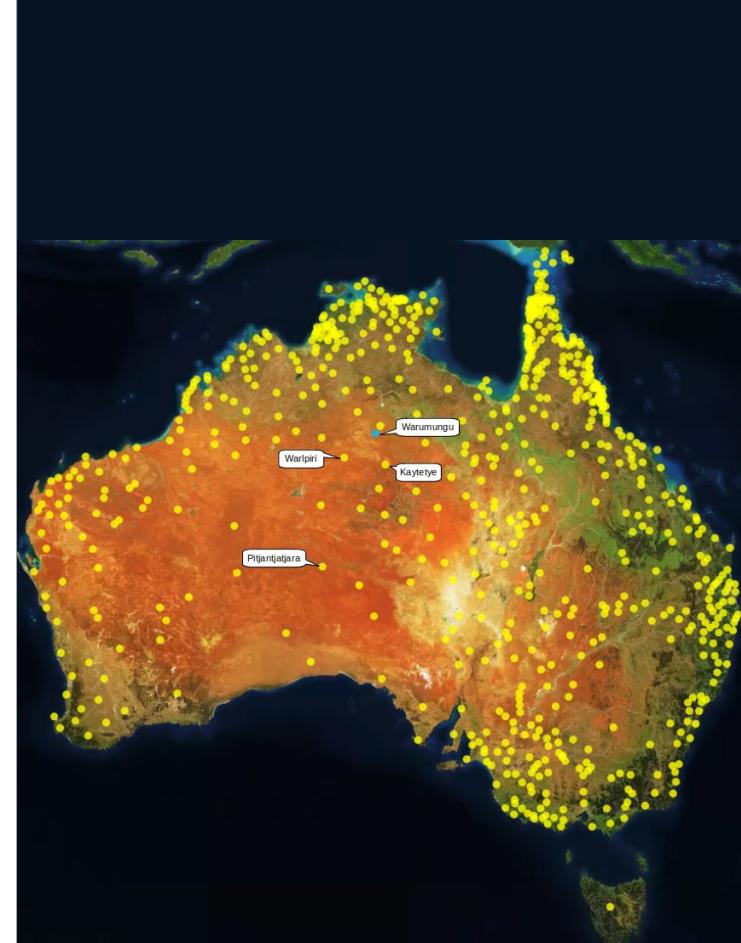


Image adapted from Gambay



# Data as decisions in a corpus

- Each point encodes a series of decisions:
  - What features to encode and how?
  - What level of detail is needed?
  - What are the relevant features for the research question?
  - How to standardize these decisions across languages & linguists?
- Each decision is like a thread that contributes to the overall pattern/picture

# Tailored corpus tool

- Addressing specific questions is often not possible with off-the-shelf tools Anthony (2012)
- Input (example):

\lx .....	headword, pos, gloss
\ps .....	Xx,yy,zz
\de .....	Aa,bb,cc
\xv .....	Dd,ee,ff
\xe .....	Gg,hh,ii



- Interactive dashboard to examine the effects of analytic decisions

Based on plotly dash



With custom python functions

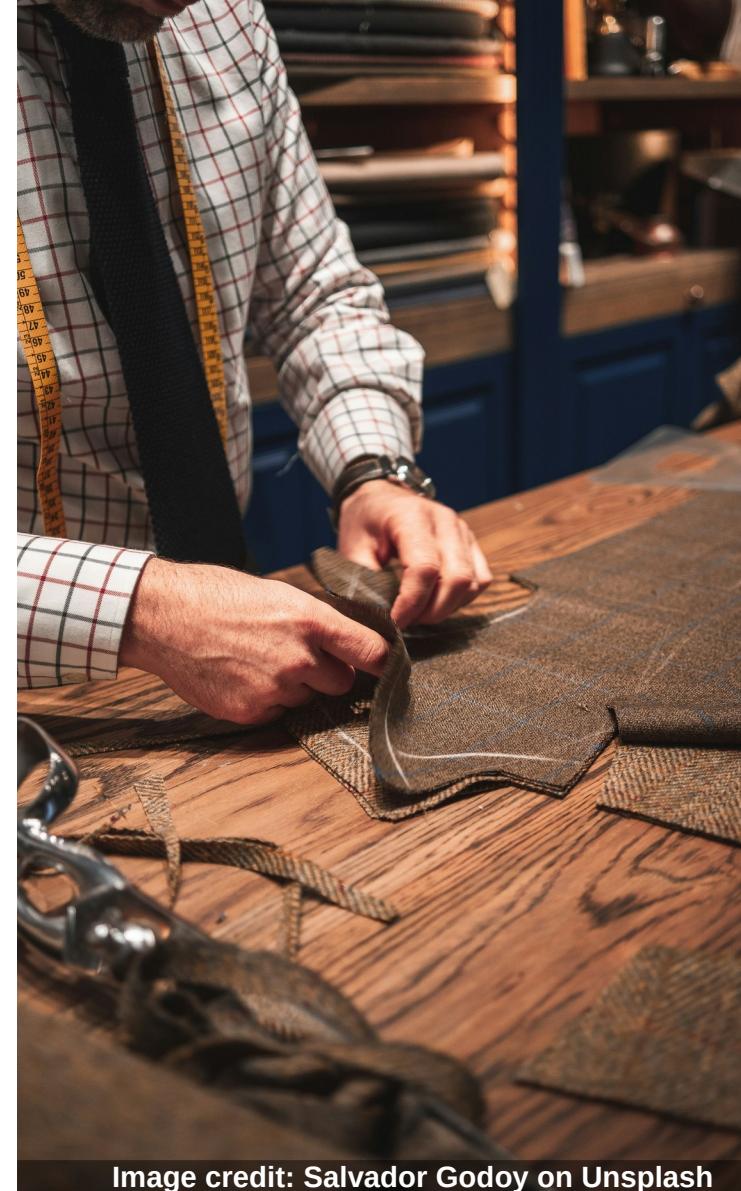


Image credit: Salvador Godoy on Unsplash

# Building Blocks Project

## Data Settings

### File Upload:

Drag & Drop or Select Files

Export

### Language:

- Kaytetye
- Pitjantjatjara
- Warlpiri
- Warumungu

### Corpus Filter Options:

- Drop duplicates
- Independent Word
- None
- Reduplication
- String removal
- Verb compound
- Verbal Morphology

### String removal Filter:

(pa)

## Analysis Settings

### Vowel Harmony Combinations:

V1V2

Vowel

Distribution

Export

### P.O.A. Combinations:

- By placement
- Consonant
- Aggregate

Tabular View

Bar Plot

File Uploaded

Name	Type	Last Modified
system-2025-02-28_13.17.01	Video	13:17
system-2025-02-28_13.15.17	Video	13:16
dash_vt.ipynb	Document	13:05
dash_vt.py	Text	3 Feb
preprocess_ipymb	Document	3 Feb
revert_ipymb.txt	Text	3 Feb
lang_operations.tsv	Text	31 Jan
phonotactic_corpora.ipymb	Text	31 Jan
legality_principle_gbb.py	Text	31 Jan
gbb_update.csv	Text	31 Jan
dash_update.ipynb	Document	31 Jan
test_table.ipymb	Document	13 Jan
test_ipymb	Text	10 Dec 2024
test_ipymb1.ipymb	Document	4 Dec 2024
dash_upload.ipynb	Document	3 Dec 2024
code_ipymb.ipynb	Text	27 Nov 2024
Analysis_dashboard.ipymb	Document	27 Oct 2024
legality_principle.ipymb	Text	18 Jun 2024
other	Text	13:14
anaconda3	Text	7 Oct 2024
test_ipymb2.ipymb	Text	10 Dec 2024
test_ipymb3.ipymb	Text	20.6 kB
dash_ipymb.ipymb	Document	3 Dec 2024
code_ipymb.ipynb	Text	4.5 kB
Analysis_dashboard.ipymb	Document	27 Oct 2024
legality_principle.ipymb	Text	6.5 kB
other	Text	6.1 kB
anaconda3	Text	13:14
test_ipymb4.ipymb	Text	27.3 kB





Image credit: James Gray

# Thank you!



Scan for repository



RESOURCEFUL-2025



Australian  
National  
University



MACQUARIE  
University  
SYDNEY AUSTRALIA



THE UNIVERSITY OF  
NEWCASTLE  
AUSTRALIA