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

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

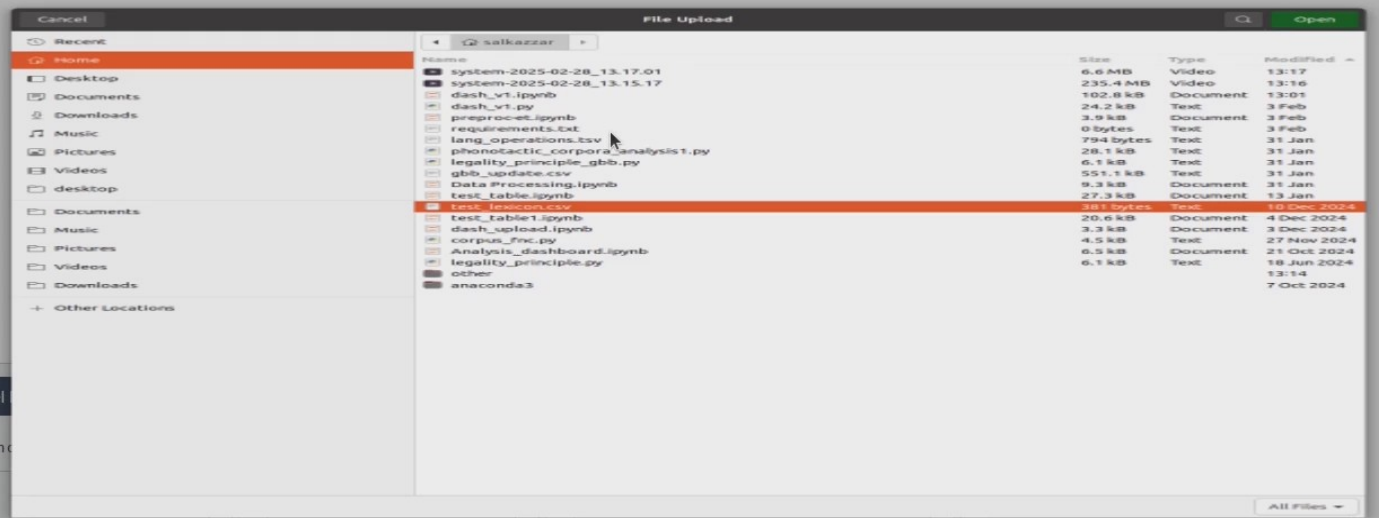
P.O.A. Combinations:

- By placement
- Consonant
- Aggregate

Tabular View

Bar Plot

Export



Vowel

Distribution

Export


Image credit: James Gray

Thank you!

Scan for repository

NODALIDA 2025 Baltic-HLT

RESOURCEFUL-2025

Australian National University

MACQUARIE University SYDNEY·AUSTRALIA

THE UNIVERSITY OF NEWCASTLE AUSTRALIA