27-29 March 2019

National Library of Australia, Canberra

Australian Academy of the Humanities' 2nd Humanities, Arts and Culture Data Summit

and

3rd international DARIAH Beyond Europe workshop

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

Australian Research Data Commons

#DARIAHBeyondEurope #HACDS2019

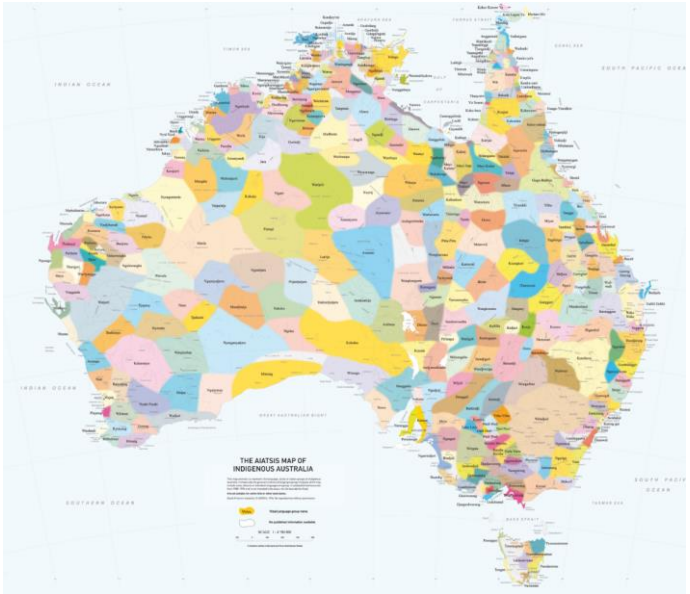
Towards an Australian Language Data Commons

Lessons from the Australian National Corpus

Michael Haugh and Simon Musgrave

Language data

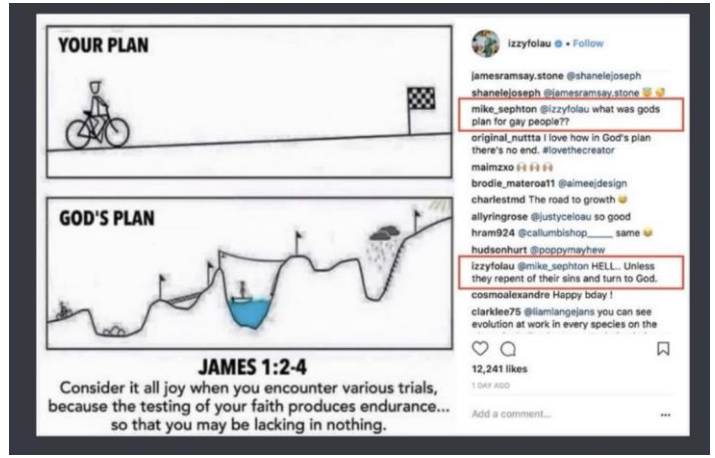
- Inclusive notion of text as spoken, written, signed, multimodal



unplanned ⟷ unplanned

institutional ⟷ non-institutional

scrabbling. 49. **(0.6)** 50. A: **hh .hh** 51. (1.0) 52. T: °like a mousemat, 129. **(1.1)** 130. A: **hhh** (0.3) no?= 131. T: =does it again fuck that 29. **(1.5)** 30. A: **hhhh** just go around 31. in circles you insane 204. **(1.0)** 205. A: **hh** 206. B: like seriously (.) like different dialects< **(0.6)** 67. A: **hh [.hh]** 68. B: [bei]ng my luck but 6. after a week **(1.4)** 7. D: **hhhh (.h) hhh:** 8. A: [of me being full of a shit 38. **(2.2)** 39. A: **hh:** (.) ↑I don't say ↓that he'll 50. D: =yeah 51. **(1.2)** 52. A: **hhhh HHHH hh (.hh)** ↑your ↓Hair ↓ney in my bank **(0.6)** 236. A: **hhh(h)** (0.3) 237. D: like (1.0) 184. A: (hm) 185. **(1.0)** 186. L: **hh .h hh** 187. A: see: I'm- <I ca 24. D: °yeah° 25. **(2.3)** 26. A: **hhhhhhhhh** 27. (1.2) 28. D: a good



Why do we need a national language data commons?

Multilingual Australia

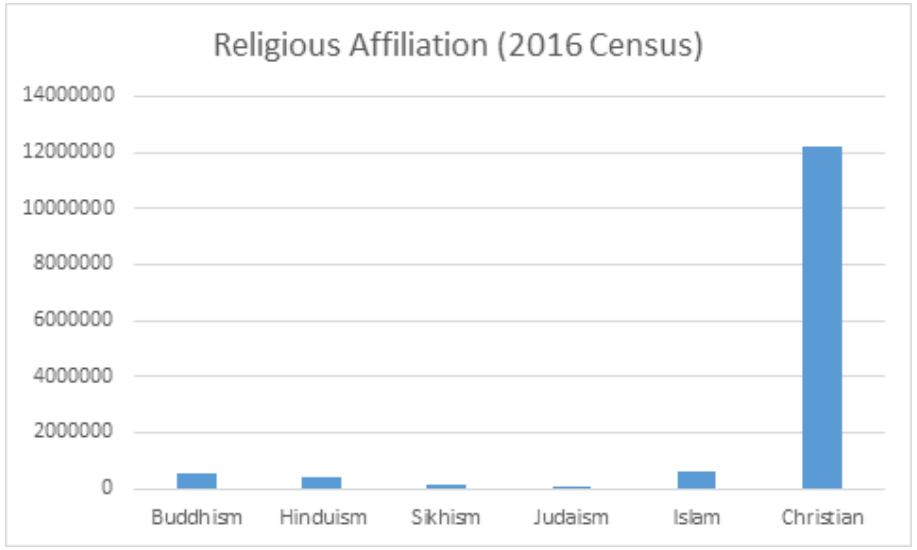
- Australia has **no** official language
- **English** is Australia's *de facto* national language (only English spoken at home by 72.7% or approx. 17 million people)
- But more than **430 different languages** are spoken in Australia
- **Australian Indigenous languages** are spoken by 0.3% of the population, or approx. 65,000 people
- 22.2% of the Australian population speaks **another language** at home

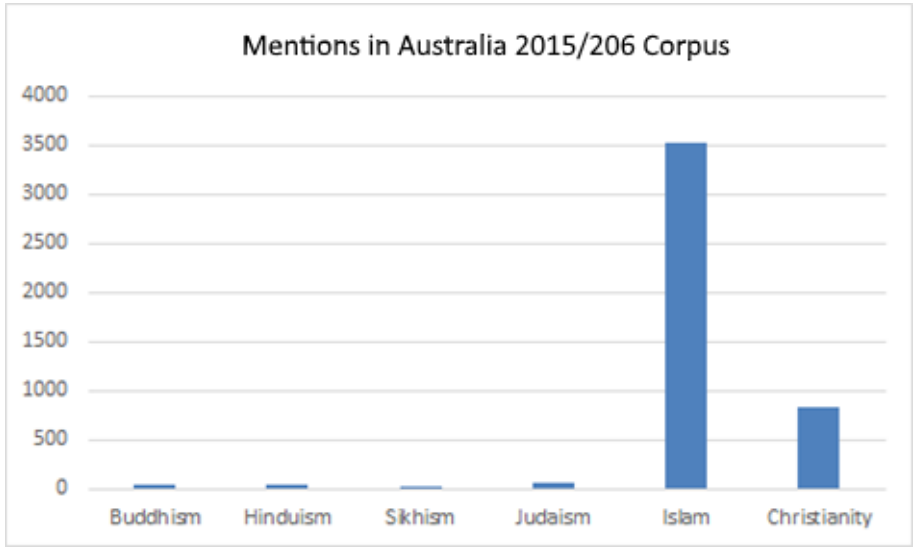
(Australian Bureau of Statistics, 2016 Census)

Social policy implications

- Social cohesion: local and national

- Tracking attitudes to race and religion over time





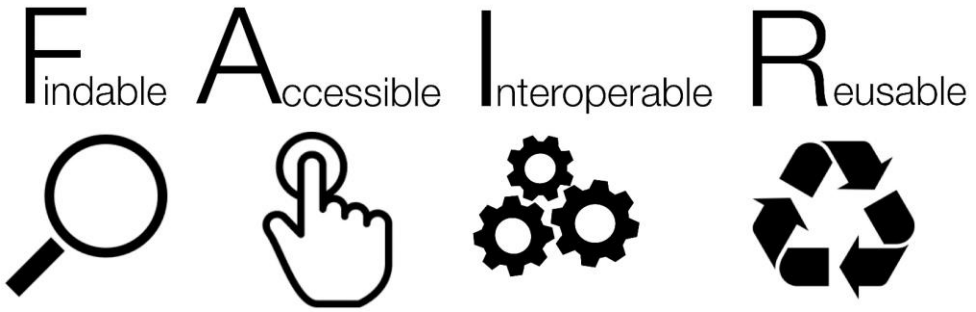
Supporting citizen science

- Language lies at the core of **cultural well-being** and Australians have an enduring interest in language

- Language **belongs to people**, to communities, to the nation

- There is a strong moral case for **democraticising** access to language data for Australians

What do researchers need?



- Language data can be infinitely **repurposed**
- Linguistics is increasingly **data driven**
- Australian researchers need language data that is **F**indable
  Accessible
  Interoperable
  Resuable

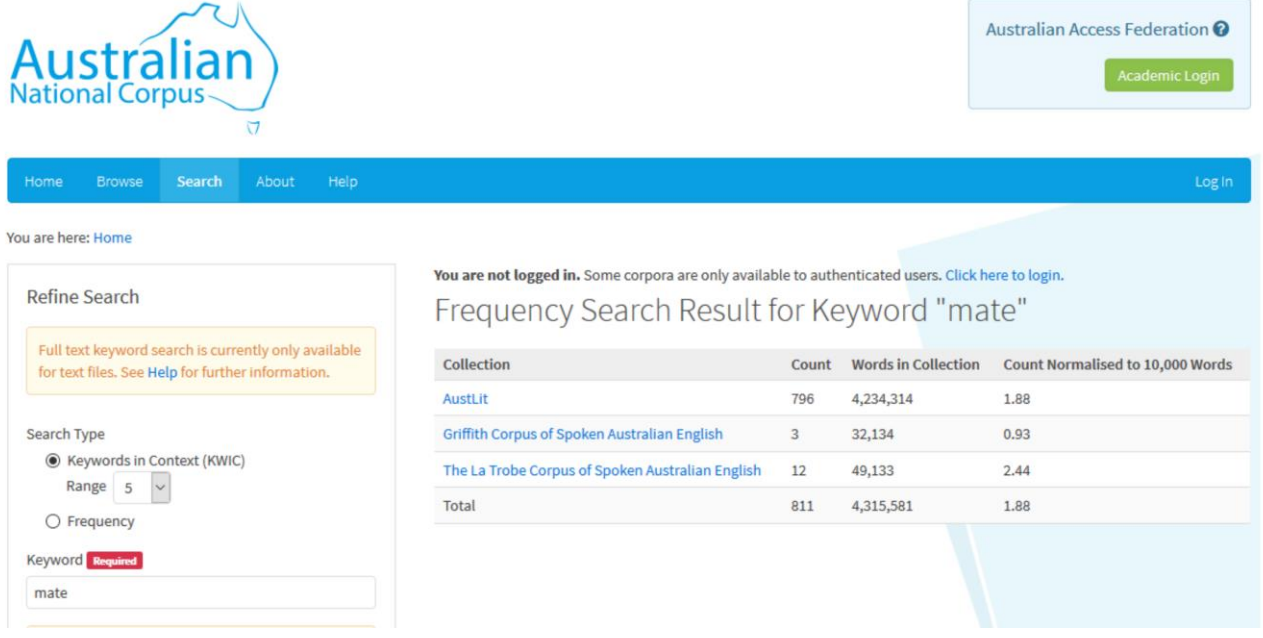
Australian National Corpus

Intentions, limitations, lessons

The Australian National Corpus initiative

- Aimed to collect and provide access to data on language in Australia

- Built on existing collections: ACE, ART, Braided Channels, COOEE, GCSAusE, ICE-AUS, MCE, LTCE (+ samples from AustLit, Mitchell & Delbridge)



Key challenges

- Location of data but also sharing data
- Building a multi-purpose interface
- Copyright and ethical problems
  – especially for legacy data

GCSAusE06 (Raw)

Item metadata

Speaker:	participant,GCSAusE06 - Participant 1,26 participant,GCSAusE06 - Participant 5,25 participant,GCSAusE06 - Participant 2,27 participant,GCSAusE06 - Participant 4,18 participant,GCSAusE06 - Participant 3,23
Contributor :	Loanne Dang
Date Transcribed :	7 October 2009
Description :	A transcribed conversation between five housemates that occurred at home. Jackson, Paul and Frank are brothers.
Ethics Approval Number :	LAL/07/HREC
Participants :	Darren (26, Australia, male, L1 English, AU, Secondary, Builder) Nate (27, Australia, male, L1 English, AU, Secondary, Crane Riggs) Jackson (23, Australia, male, L1 English, AU, Secondary, Form Worker) Paul (18, Australia, male, L1 English, AU, Secondary, Form Worker) Frank (25, Australia, male, L1 English, AU, Secondary, Apprentice Electrician)
Transcribers :	Loanne Dang (April 2009)
Audience :	Small Group

File contents

```
Transcript Coversheet
|                                |Data
|
|Title                           |GCSAusE06
...|
|Number of people                |5
|
|Description                     |A transcribed conversation between five housemates that occurred
at     |
|                                |home. Jackson, Paul and Frank are brothers.
|
|Participants                    |Darren (26, Australia, male, L1 English, AU, Secondary, Builder)
|
|                                |Nate (27, Australia, male, L1 English, AU, Secondary, Crane Riggs)
|
|                                |Jackson (23, Australia, male, L1 English, AU, Secondary, Form
Worker)   |
|                                |Paul (18, Australia, male, L1 English, AU, Secondary, Form Worker)
|
|                                |Frank (25, Australia, male, L1 English, AU, Secondary, Apprentice
|
|                                |Electrician)
|
```

(Unintended) consequences of decisions

- Homogenising data (multiple versions of data)
- Location of people, relation to funding sources, tie to one institution all become problematic
- Version control – sharing with Alveo
- Constructing the platform was a short-term project: whose responsibility is it to add new data?

Towards an Australian Language Data Commons

Federating Australia's language data

- **One quarter** of the **world's languages** are spoken in the Pacific South West – many of these are **endangered** languages

- Australian researchers have collaborated on **curating data** on many of these languages: (1) Australian indigenous languages, (2) indigenous languages in Australia's region, (3) Australian English(es), (4) Australian community languages

PARADISEC

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE

Australian
National Corpus

Things to do differently

- Sustainable **governance** is important: independent from individual institutions (and personalities) as far as possible
- Equitable and ethical **access** to language data is the primary goal – tools for analysing language data remain separate
- Ensuring adherence to **metadata standards** more important than imposing data standards as language data can be highly granular
- Representativeness is not a goal, but possibility of building **representative sub-collections** (in various ways, for various purposes)

Expanding the circle

- Australian Language Data Commons as a node of a HASS Data Commons
- Currently in consultation with AIATSIS

- Trove and Pandora as (largely untapped) resources for linguistic research
- Australian broadcasters…

Expanding research strategies

- Aggregation of data is a **cost-effective** way to enable access to large data sets
- But then we are challenged to find ways to **exploit** such resources
- Example: combinatorial search across collections to assemble candidate sets of examples (Haugh & Musgrave 2018)
- Working with a language data commons opens new possibilities:
  - New research questions
  - New research methods