

---

RILCA, Mahidol University Seminar 2 May 2024

# Working with legacy materials in language documentation and revitalisation: challenges and opportunities

Peter K. Austin

Department of Linguistics  
SOAS, University of London

---

---

© Peter K. Austin 2024

Creative commons licence

Attribution-NonCommercial-NoDerivs

CC BY-NC-ND

Thanks: Andy Cowan, Andrew Garrett, Ed Garrett, Philip Jones, Jorge Labrada, David Nash, David Nathan, Jane Simpson, Clara Stockigt and Nick Thieberger,

Some of the research reported here was supported by a Leverhulme Emeritus Grant (2021-2023)

---

---

# Overview

- Goals of language documentation, description, revitalization
  - What are legacy materials?
  - Challenges
  - Opportunities
  - Two case studies:
    - Guwamu, Southern Queensland – S.A. Wurm's fieldnotes
    - Diyari, South Australia – J.G. Reuther's dictionary
  - Discussion and conclusions
-

---

# Language description

- Study of language as a **system** separated from its actual use by speakers and the social-political-cultural-economic conditions of use
  - Requires abstraction and search for general principles (phonology, morphology, syntax, semantics, pragmatics)
  - Requires idealisation and “cleaning up” recordings of actual use
  - Data collection often involves **elicitation** through surveys or interviews or **experiments**
  - Studying a language the researcher does not speak is often done via **translation** or asking for **speaker judgements**
  - The records of interview or survey are **not** of interest in themselves, but just a way to accumulate “**the data**” for analysis
-

---

# Language description

## 1. **Goals** of description:

- to present language structures for others to understand;
- to identify common features and differences across languages (typology);
- to understand how the human mind works (psychology, neurophysiology);
- to understand how humans interact and express personal, social and cultural relationships

## 2. Analysis is often highly structured and formal and written in an abstract **metalanguage**

## 3. the **audience** for description is typically other researchers, and distributed in books or articles (grammars, dictionaries, maps, graphs, narratives, text collections)

---

---

# Language documentation

- “concerned with the **methods, tools, and theoretical underpinnings** for compiling a **representative and lasting multipurpose** record of a natural language or one of its varieties” (Himmelman 1998)
  - Features:
    - *Focus on primary data*
    - *Accountability*
    - *Long-term storage and preservation of primary data (archiving)*
    - *Interdisciplinary teams*
    - *Cooperation with and direct involvement of the speech community*
  - Outcome is **annotated and translated corpus** of archived representative materials on a language, cf. TLA/Dobes, ELAR
-

# Language revitalisation

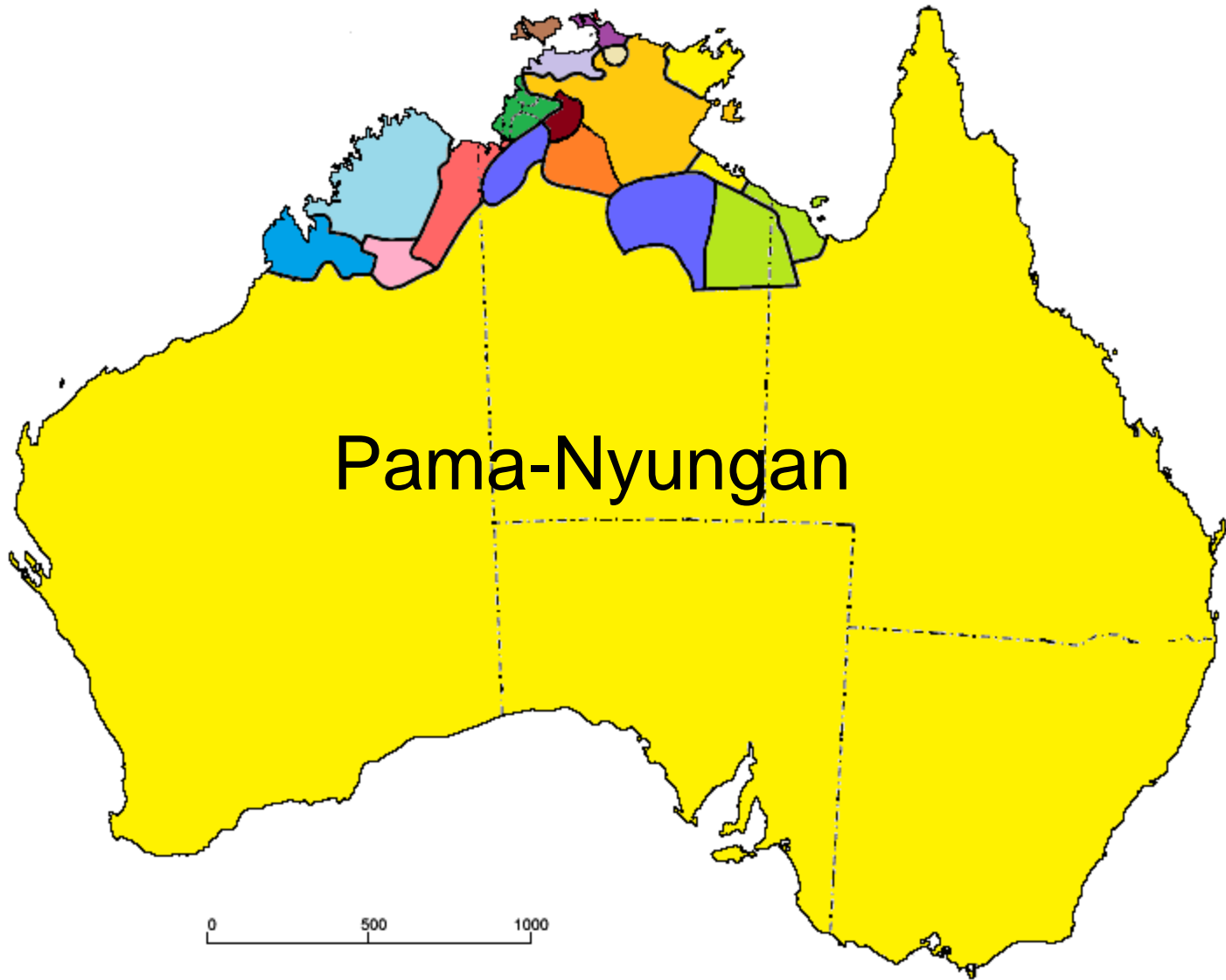
- efforts to increase **language vitality** by taking action to:
  - increase the domains of use of a language and/or
  - increase the number of speakers (often in the context of reversing language shift) both adults and children
- older than language documentation (serious work began in 1970s and 1980s among Maori, Native American groups and others)
- speech/language community members are often more interested in revitalisation than documentation
- often assumed revitalisation = formal language learning (school lessons, immersion)
- many communities are now using corpora to support language learning

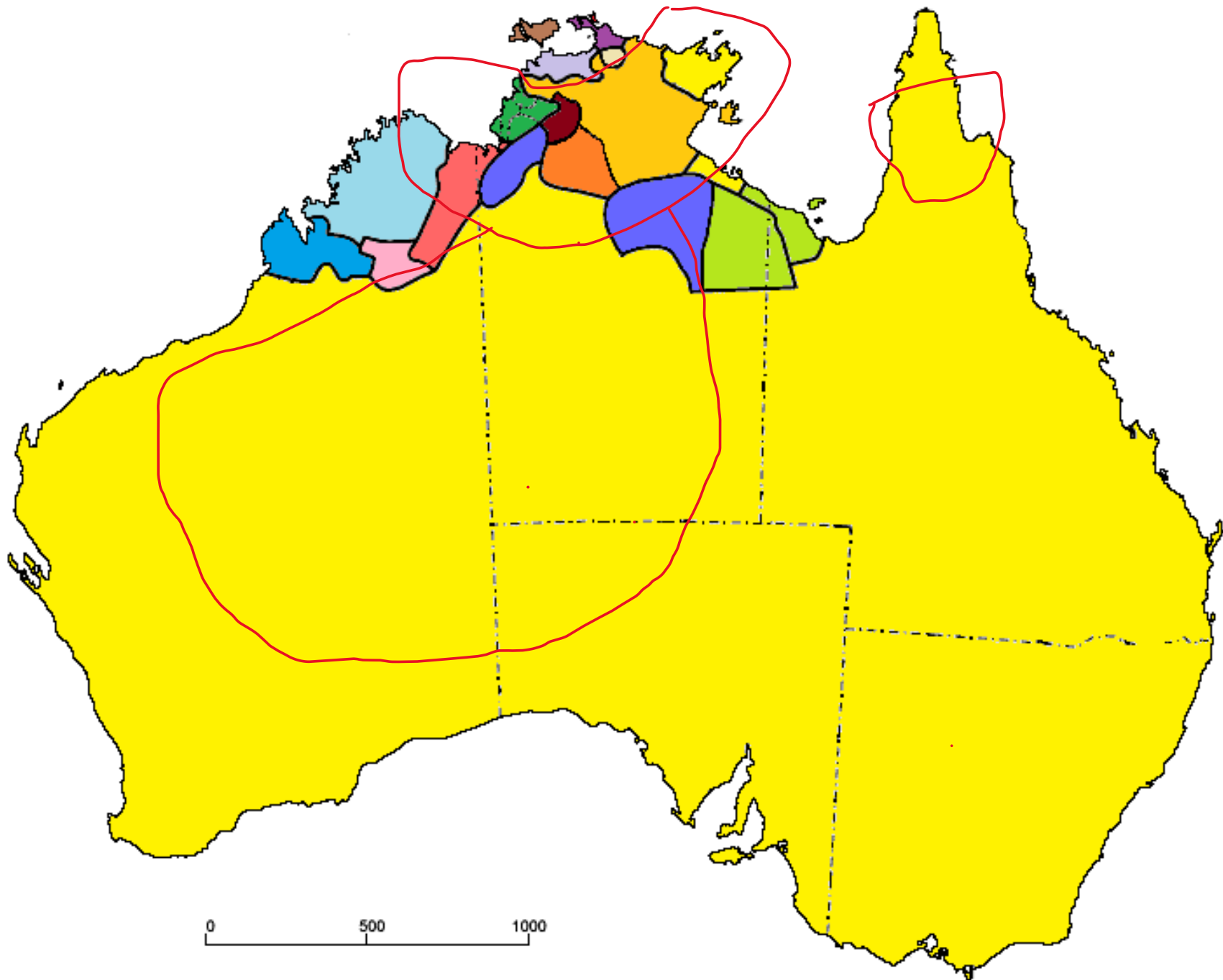
---

# Legacy materials (secondary corpora)

- It is rarely the case that first-hand research is carried out on languages or communities that have **never** been documented before, so typically there already exists material in some form, in missionary or traveller reports, government records, or from previous linguistic or anthropological researchers. With careful use, these **legacy materials** can provide valuable information to contemporary researchers and communities, and assist language recovery or revitalisation
  - In some cases there are **no** contemporary fluent speakers and legacy materials are the richest or only sources for description and revitalisation
  - Sometimes, field research in communities is not possible due to danger from violence, e.g. civil war or gangs, or from disease, including pandemics like Ebola and Covid-19. Legacy materials can be important in such cases.
-







---

# Corpora epistemology issues

- **biography** of creator(s): prior language knowledge and/or study and/or exposure, their teachers/mentors/correspondents, how/when they learnt the language, how long they worked on the language and at what point in their careers, how the work was funded and with what goals, whether there were previous studies of the language or the community that they could have had access to, who was the consultant, what prior experience, who did the value-adding (transcription, translation)
  - aspects of **historical period**: kind and impact of contact between communities, including colonialists, and influential descriptive categories and formats known to author(s), e.g. traditional grammar based on Latin or Greek models, structural linguistics, IPA
-

---

# Form, content, and interpretation issues

- Audio may be poor quality, noisy, difficult to hear (e.g. multiparty conversations)
  - Video may be poorly recorded or unwatchable, poor audio, partially record people out of frame
  - Hand-written or typed text can be difficult to read or interpret, with crossing out, abbreviations, obscurities (requires philological analysis)
  - In digital text files characters may be mismapped or missing due to font problems, tabbed text may not align, structured text may be uninterpretable if the structure definition is missing
  - Retranscriptions should link back to documents or files on which they are based (so we can retrace the steps of analysis and value-adding)
-

# Retranscription example



## Book B, Page 8

◀ Prev

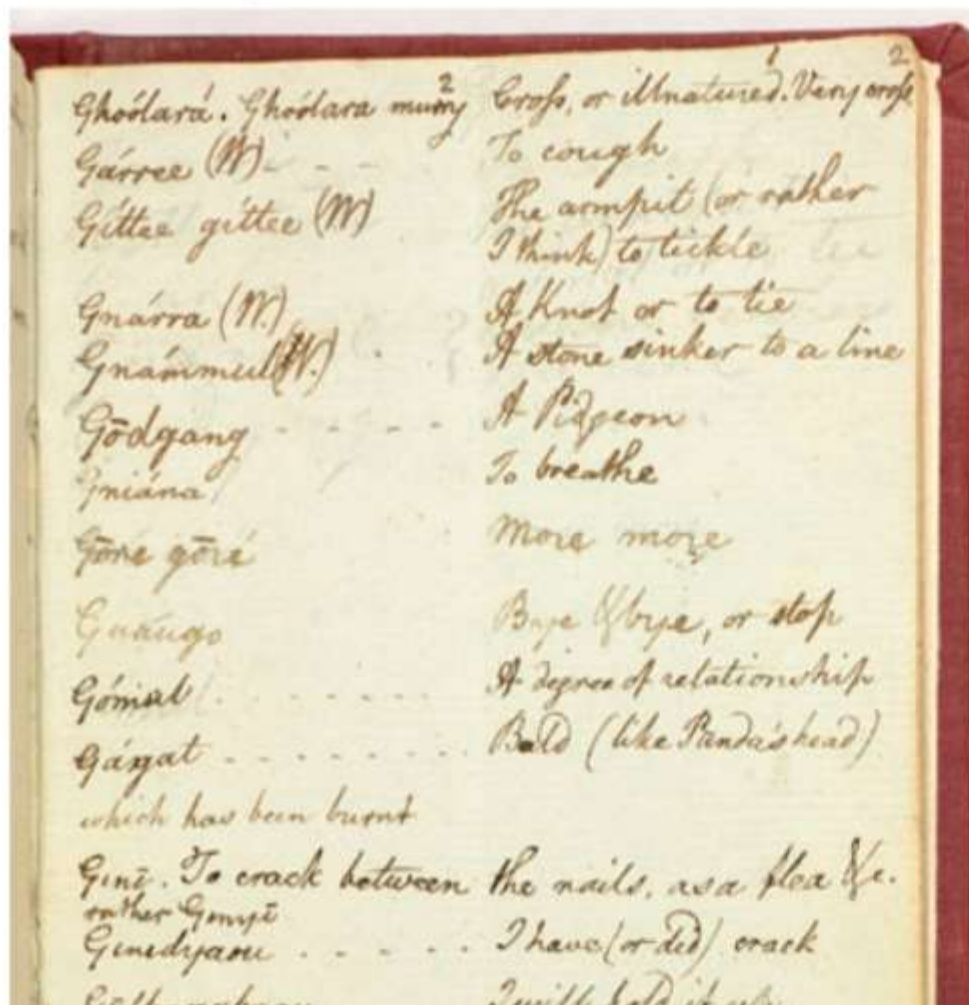
Contents ▲

Next ▶

Manuscripts Book A ▾

Manuscripts Book B ▾

Manuscripts Book C ▾



Ghoólara <sup>1</sup> . Ghoólara murry <sup>2</sup>	Cross, or illnated <sup>1</sup> . Very cross <sup>2</sup>
Gárree (W)	To cough
Gíttee gíttee (W)	The armpit (or rather I think) to tickle
Gnáarra (W.)	A knot or to tie
Gnámmul (W.)	A stone sinker to a line
Gödgang	A Pidgeon
Gniána	To breathe
Göre göré	More more
Guáugo	Bye & bye, or stop
Gómül	A degree of relationship
Gáyat	Bald (like Púnda's head)
which has been burnt	
Gini. To crack between the nails, as a flea &c.	
rather Ginyi	I have (or did) crack
Gülbarabaou	I will hold it up
Gonajülye desiring to wear one of Patyegaran's petticoats. I told her it was too long for her, on which she said Gülbarabaou which Patye explained as above. -	
Gwára buráwá	The wind is fallen.



Book B, Page 8

◀ Prev

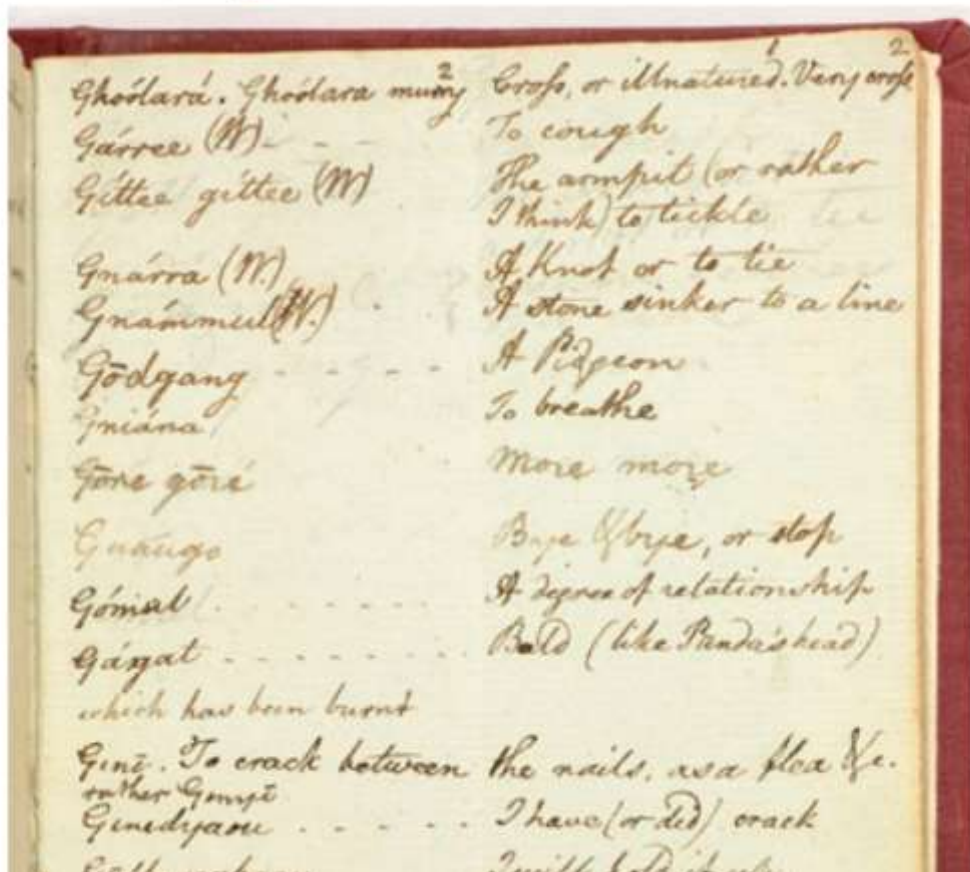
Contents ▲

Next ▶

Manuscripts Book A ▼

Manuscripts Book B ▼

Manuscripts Book C ▼



Ghoólara<sup>1</sup> Ghoólara murry<sup>2</sup>

Gáree (W)

Gíttee gíttee (W)

Gnáarra (W)

Gnáimmul (W)

Gódgang

Gniána

Góre góre

Guáugo

Gómul

Gárgat

which has been burnt

Gini. rather Ginyi To crack between the nails, as a flea etc.

Giniyaou

Gálbarabaou

Gonaúlye desiring to wear one of Patyegarar's pettycoats: I told her it was too long for her, on which she said Gálbarabaou which Patye explained as above. -

Gwara buráwa

Cross, or illnatured<sup>1</sup>. Very cross<sup>2</sup>.

To cough

The armpit (or rather I think) to tickle

A knot or to tie

A stone sinker to a line

A Pidgeon

To breathe

More more

Bye & bye, or stop

A degree of relationship

Bald (like Púnda's head)

I have (or did) crack between the nails

I will hold it up

The wind is fallen

# Models for representation of structured text

- *Relational model* – identify **entities** and **relationships**, typically encoded via **database** software, e.g. form, gloss in one-to-many relationship

ID	Form	Gloss
001	marda	stone
002	marda	money

- *Markup model* – hierarchical representation that uses tags (XML) and scoping to encode entities and dependency relationships, e.g. `<entry><form>marda</form><gloss seq="1">stone</gloss><gloss seq="2">money</gloss></entry>`



---

# Form, content, and interpretation issues

- implicitly structured materials, e.g. use typography or page layout to distinguish analytical categories or kinds of information, can be made more useful by encoding the structure separately from the form, e.g. by XML markup + stylesheet, or database model + display process
  - structure is not always computable from typography and may need to be manually added (cf. quotes from Nathan in Austin 2017): over-use of quotation marks, for multiple purposes and often redundantly, unclear scoping, spelling errors
  - cryptic or incorrect glossing, because author(s) could not understand language consultant's accent or pronunciation, or because the semantics of the source language terms were misunderstood (the "gavagai problem")
  - changing interpretations over time, and the author(s) hearing what they think they heard rather than what is in the recording or was dictated by speakers due to analytical decisions
-



Table 1 Semantic problems in text materials, from Crowley and Austin (2005).

	Wordlist meaning	Correct meaning
1.	Pronunciation problems	
	heart	hot
	wet	sweat
	moths	boss
	dung, shit	tongue
2.	Meaning problems	
a.	generic versus specific	
	grass	vegetation
	boy	uninitiated youth
	beard	hair
	day	now
	thumb	your hand
	girl	female
b.	related word	
	thighs	buttocks
	cloud	sky
	woman	wife
	hair	head
	frown	blind
	spider	to bite
	dig	drink

---

# Form, content, and interpretation issues

- Inappropriate content (taboo, sacred) for various audiences
  - Dated content using expressions that are no longer acceptable or now inappropriate, e.g. personal remarks about the ancestors of living persons
  - Over-distinguishing or under-distinguishing crucial contrasts, in phonology (voicing, aspiration, vowel quality or quantity, tone), morphology (ergativity), or syntax (vocative case, applicative, switch-reference)
  - Lack of sociolinguistic context: who says what to whom when and where?
  - Relationship between corpus forms and contemporary usage/beliefs – issues about what is “right”, especially for shifting languages undergoing change
-

---

# Stakeholder issues

- projects typically have many stakeholders who may have different kinds of interests in the materials collected and the analyses created
  - Issues of control, consultation, and decision-making are important when deciding what kind of documentary material to include in any corpus and how it can be used
  - For legacy materials possible mismatches between past situations and the present
    - current membership of a contemporary ‘community’ may not coincide with past membership
    - people who provided legacy materials may not even now be viewed as rightful members of a given group and therefore their information may be deprecated
-

---

# Stakeholder issues

- Unclear agreements, if any, between original collector and the community or particular individuals at the time (and whether these agreements were documented) as well as the relationship between any such agreements and arrangements that are currently being negotiated between contemporary researchers and other stakeholders, e.g. Austin told not to distribute copies of Wurm's materials without permission of current Aboriginal group who self-identify as descendants
  - Best to clarify if possible before creating and using the corpus
-

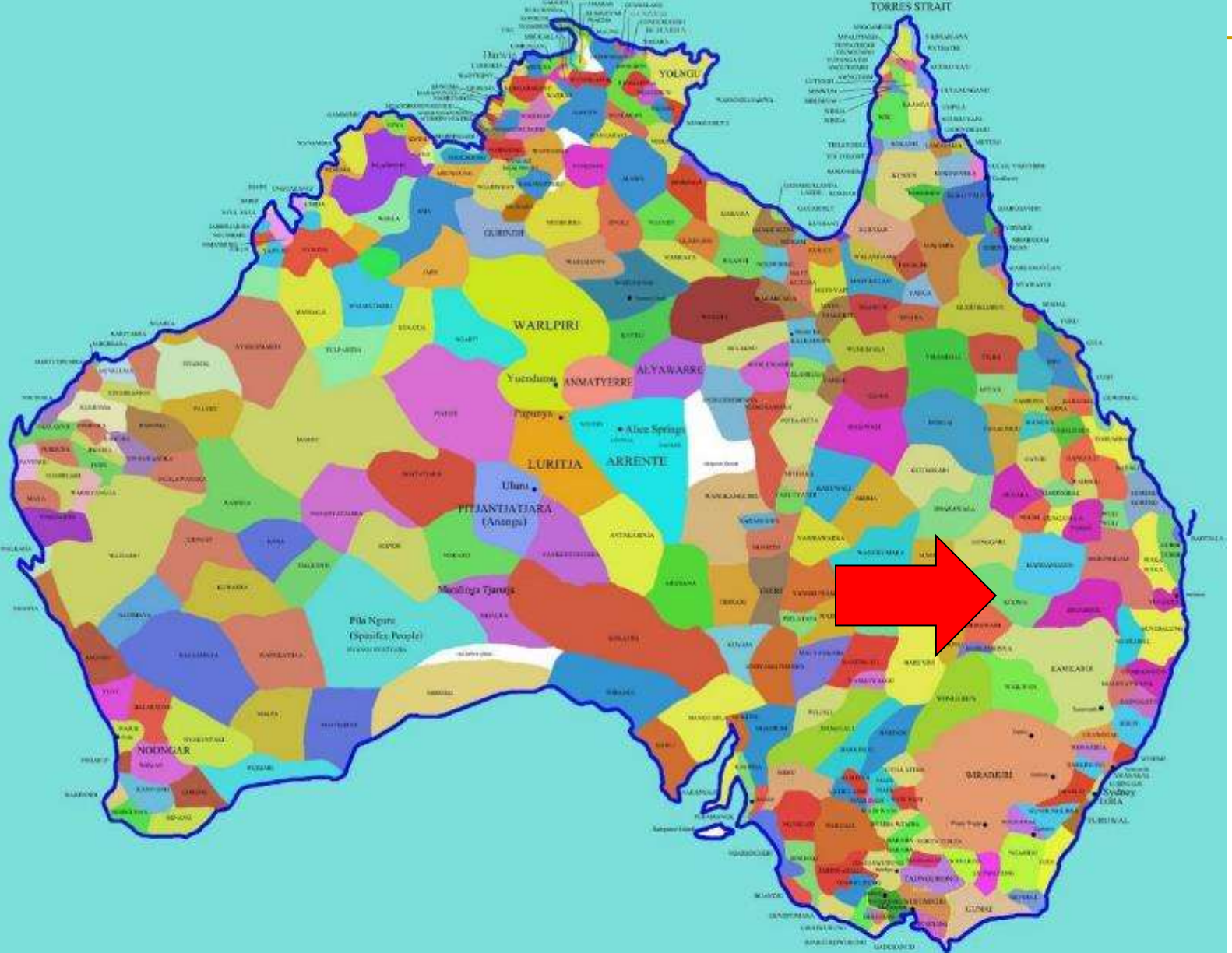
# Rights issues

- who holds what rights? Are the rights documented? How do we establish rights retroactively? What if the researcher is not sure about speaker rights?
- how do we determine rights when there are multiple contributors and data comes from multiple media?
- who has inherited rights between the time of the original recording and now? (e.g. descendants of the original speakers, descendants of the original researcher)
- what happens to 'orphan works' where the original stakeholders can no longer be identified? (e.g. materials passed from a researcher to a later researcher)
- when analysing corpus data it is important to clearly document the various contributions to the work, including those of the original author(s), research assistants, the linguist-editor, other researchers, and current community members.
- access rights to secondary corpora need to be decided and clearly documented

---

# Legacy materials case study 1

- Survey fieldwork by S.A. Wurm 1955-57, multiple languages from eastern Australia, materials collected but not analysed
  - For Guwamu (Kooma), southern Queensland we have 13 minute tape recording plus 60 pages of fieldnotes
  - Phonetic transcription of elicited words and sentences
  - Glosses in Hungarian shorthand, verbally translated onto tape by Wurm 1978
  - No contemporary speakers
-





(Epi. Taylor) (Mac Taylor) (R) (M.O.) daddja layambuni  
wanju inipi punju - ε

ε'ε - εε, (ε ε ε / ε ε ε ε, - 'ε ε ε)

ε'ε, wala do daddja layambuni inipi punju

minju inipi punju ε'ε

banimaina leu hui daddja lambunna  
wanitma jagu! unu

buppani punju ε ε ε ε ε ε ε ε ε ε

punju ε

inipida jalida (punju ε ε ε ε ε ε ε ε ε ε)



---

# Guwamu case study

- Retype original notes, add consistent practical orthography, add metadata on sources (speaker, recorder, fieldnotes location), abbreviation definitions, date of last edit
  - create **lexicon**: headword, gloss, definition, scientific name, scientific name source, picture, semantic relations (synonym, antonym, cf), notes, cognates, example sentence link (text, free translation)
  - create **sentence analysis**: phonemicization, morpheme glossing, part of speech, free translation in English, notes, link to lexicon (lexnum), link to abbreviations
-

Toolbox - Guwamu\_lexicon.bt

File Edit Database Project Tools Checks View Window Help

[no filter]

Guwamu\_lexicon.bt

lx	<b>badha</b>
la	
lu	
lxnum	001
lg	Gu
lps	v
lsub	vtr
lge	bite
lde	bite
lsd	Actions
lsci	
lsci-source	
lnt	
lsyn	
lant	
...	

guwamu\_abbreviations.bt

labb	<b>vtr</b>
lmng	transitive verb
ltype	<b>sub-category</b>
lnt	transitive verbs are a sub-category of verbs; they take a transitive subject argument in ergative case and a transitive object argument in accusative or absolutive case.
lcf	<b>vi, vdi</b>
lgr	
ldate	03/Apr/2005

Guwamu\_FellData.bt

lref	Guwamu.001
lwmm	baðalgula jinunna
lx-A	badhalgula yinunha
lm-A	badha -lgu -la yinu -nha
lmng-A	bite -fut -3sg 2sgdat -acc
lcat	v -suff -suff pro -suff
lsubcat	vtr -vinfl -proagr pro -proinfl
lxnum	001 -012 -011 077 -024
lg	It will bite you.
lnt-B	
lnt-A	
lrec	SW
lsp	WW
lcreator	JB

guwamu\_people.bt

lid	<b>WW</b>
lname	Willy Willis
lrole	<b>speaker</b>
lg	<b>Guwamu</b>
lnote	speaker of Guwamu language interviewed by SAW
lcf	<b>SAW</b>
ldate	06/Jul/2023

```

lx      badha
la
lu
lxnum  001
lg      Gu
lps     v
lsub    vtr
lge     bite
lde     bite
lsd     Actions
lsci
lsci-source
lnt
lsyn
lant
lcf
lder
lmargany badha
lgunya   badha
lbidjara badha
lmuruwari
lrec     SW, LO
lsp      WW
lref     Guwamu.001
lv       badhalgula yinunha
lvge     It will bite you.
lref     Guwamu.220
lv       Gundu wardayanda ngurrangundu badhandula yinha
lvge     Go away from the dog, it may bite you.

lwnum
ldate-entry 13/Mar/2005
ldate-edit  01/May/2024
    
```

```

lref     Guwamu.001
lwmm     badhalgula jinunha

ltx-A    badhalgula yinunha
lm-A     badha -lgu -la yinu -nha
lmg-A    bite -fut -3sg 2sgdat -acc
lcat     v -suff -suff pro -suff
lsubcat  vtr -vinfl -proagr pro -proinfl
lxnum    001 -012 -011 077 -024

lg       It will bite you.
lnt-B
lnt-A
lrec     SW
lsp      WW
lcreator JB
lsource  Wm-p01Bs09
ldate    06/Jul/2023
    
```

---

# Outcomes to date

- Guwamu-English reference dictionary
  - Guwamu-English learner's dictionary
  - Guwamu children's story by Cheryl Levy and Christopher Bassi
-

banggard

**banggard** *n.* back. *Category:* Body part.  
*Etm:* Guuya, Margany.

**banggu** *n.* hard dirt, stone.  
*Category:* Nature. **Dhambal**  
**banggunggu gudjayanda gardarru**  
**gundu** Hit the snake with a stone before it  
gets away. **Marda waayanda**  
**ngadjunha banggu banyarri**  
**yilunha dhabalguli** Give me a hand, we  
two will roll this big stone.

**banggun** *n.* head. *Category:* Body part.  
*Etm:* Guuya.

**banyarri** *n.* big. *Category:* Qualities.  
*Ant:* gayugaanba. *Etm:* Guuya.

**banydja** *vt.* to sing. *Category:* Language.  
*Note:* southern Guwamu dialect  
*Syn:* wumba. *See:* mudhun. *Etm:*  
Guuya, Margany. **Muginydu mudhun**  
**banydjalgula** The woman will sing.

**banydjurd** *n.* belly, stomach.  
*Category:* Body part. *Etm:* Bidjara,  
Guuya, Margany.

**bara** *vt.* to hit with hand.  
*Category:* Actions.

**bardal** *adv.* tomorrow. *Category:* Time.

**bardi** *adv.* maybe. *Category:* Particle.  
*Etm:* Margany.

**bardu** *n.* river. *Category:* Nature. *Etm:*  
Guuya, Margany.

**-bari** *nder.* having nominal derivational  
affix allomorph. *Category:* Grammar.  
*Note:* used after consonant-final stems  
*Syn:* -wari.

**barra** *vt.* to scratch. *Category:* Actions.  
*Syn:* baa. **Barrayanda nganha**  
**banggarda** Scratch me on the back!  
**barra-la** *vi.* to scratch oneself. **Bawurra**  
**barralanbangala** The red kangaroo is  
scratching itself.

**barran.girri** *n.* vulgar, coarse.  
*Category:* Qualities.

**barrangirri** *quant.* one.  
*Category:* Quantities. *See:* bulardi,  
gagurbarri.

bii

**barrarru** *adv.* quietly. *Category:* Qualities.  
*Ant:* mirra.

**barrbirta** *n.* echidna. *Category:* Animals.  
*Tachyglossus aculeatus.*



**barrbu** *vt.* to spear. *Category:* Actions.

**barri** *vi.* to cry, weep. *Category:* Language.  
*Etm:* Bidjara.

**barriny** *n.* burnt. *Category:* Qualities.

**barrugirri** *n.* name of Southern dialect,  
spoken by Willy Willis.  
*Category:* Language.

**bawul** *n.* chicken, fowl. *Category:* Birds.  
*Note:* loan from English

**bawurra** *n.* male red kangaroo.  
*Category:* Animals. *Ocphranter rufus.*  
*Note:* used as a generic term for kangaroos  
*See:* gula; gumbarr; dhugandu.



**bayuli** *vi.* to crawl. *Category:* Motion.  
**Ngunan dhambal bayulinydjangala**  
A snake is crawling along there.

**-bi** *suff.* plural? *Category:* Grammar.

**bidhal** *n.* bark of tree. *Category:* Trees.  
*Etm:* Bidjara.

**bidhu** *n.* other, other one.  
*Category:* Quantities. **gandu gaba**  
**yibanydjangala yilunggu bidhunggu**  
**bundhalala nhunga gardarru gamu**  
**ganydjagirri** The child really drowned.  
That other one (child) picked it up, before  
the water would have choked him.

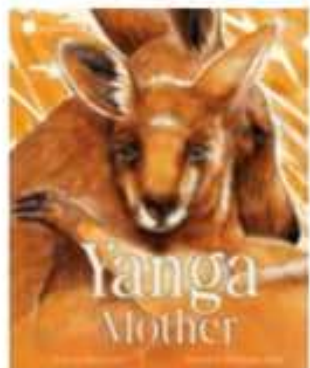
**bigany** *n.* fingernail. *Category:* Body part.  
*Etm:* Muruwari.

**bii** *adv.* can. *Category:* Particle.



FIRST  
NATIONS

BILINGUAL



## Cheryl Leavy and Christopher Bassi *Yanga Mother*

*Yanga Mother* is a timely and poetic celebration of First Nations languages. This powerful bilingual story honours connection to Country and the unbreakable bonds of never-ending motherly love.

*Händaguli Yanga. There is always Mother.*

From award-winning writer Cheryl Leavy comes this beautiful picture book in Kooma and English about a grey kangaroo and her joey, and the unbreakable bonds of family.

With artwork from renowned Meriam and Yupungathi artist Christopher Bassi, this gentle yet powerful story honours the Stolen Generations, First Nations matriarchs, and never-ending motherly love.

ISBN 9780 7022 8831 1

ANZ release July 2024

RRP AUS\$28.99

Category Picture book

Format Hardback | 316 x 250mm | full colour

Extent 32pp

Rights held World Inc. film/TV

*"Yanga Mother is a beautiful, heart-warming story that celebrates the power of Language to comfort young readers and teach them the importance of Country and how it nurtures us always."*

— Yasmin Smith, Editor



**Cheryl Leavy** is an award-winning writer and poet from the Kooma and Ngugi Nations in western and central Queensland. She is passionate about language revitalisation and working in her own language. In 2022, Cheryl won the Gudgeroo Nonclassical Poetry Prize and has appeared at Brisbane Writers Festival, Byron Writers Festival, with the Queensland's Chamber Orchestra, and more. She has served on several boards including the Queensland Art Gallery, Canberra Museum and Gallery, ACT Cultural Council, the Queensland Music Festival, and currently the Brisbane Writers Festival and the Institute of Modern Art.

**Christopher Bassi** is an artist of Meriam, Yupungathi and British descent. Working with archetypal motifs of representational painting, his work engages with the medium as sociological and historical text and as a means to address issues surrounding cultural identity, alternative genealogies, and colonial legacies in Australia and the South Pacific. Chris's recent work will be included in the Museum of Contemporary Art's exhibition, *Primavera 2023: Young Australian Artists*.



Ylungga yabangga  
Everywhere

Wandhandja-wandhandja  
Everywhen

Page Proofs | Final pages from *Yanga Mother* by Cheryl Leavy and Christopher Bassi



Yanga dharudu

Mother is sun

---

## Case Study 2

- Bilingual dictionary of the Diari/Dieri/Diyari language created by missionary Rev. J.G. Reuther
  - Head of Killalpaninna Lutheran Mission 1888-1906
  - 1897 co-authored translation of Bible New Testament (first in Australia)
  - Compiled 14 volume manuscript of Diyarri language and culture, volumes I to IV are a Diyarri-German dictionary
  - Purchased by South Australian Museum 1917
  - Rev. P. Scherer translates to English 1974
-

# Diyari language and Dieri people





# J.G. and Pauline Reuther





# Scherer Diari-English translation

1885

IV,80 9) kumari tapana = 'to drink blood; to suck at a wound'

When he is first being made a man, the wiljaru opens his mouth widely and drinks the blood [offered to him],

10) kumari tapana = 'to drink blood', i.e. to drink the blood that has been washed off a spear, wherewith a man has been killed. The young men have to do this, in order to become fearless,

11) danju tapana = 'to drink danju 'a variety of fruit'. In this instance one does not [use the term] 'eat',

12) paljangani tapana = 'to chew<sup>1</sup> a variety of [tree] gum'

IV,81 13) kirra tapana = 'to inspect a boomerang time and again' (during the making, to see if it is straight).

14) kana tapana = 'to devour the people.' The kutji does this.

15) tapana is used for 'eating' of the following fruits: kudnampirra, mpiampia, narimai, ngaliaru [and] nguratikiri.

16) mara tapana = 'to join in the eating of seed' which belongs to someone else.

17) wona marujeli tapana = 'for an old wona 'digging stick' to absorb' the water. This is placed in the water and enchanted, so that it may absorb the water,



---

# Diari-English dictionary: some statistics

- 2,180 page typescript (published as microfiche)
  - Digitisation and XML markup (Austin 2023): 4,262 entries, 15,955 sub-entries, 27,472 Diyari expressions, 13,133 (classified) notes, 3,879 examples (Diyari, literal translation, free translation), 1,766 translator footnotes, 1,273 German expressions
  - Main dictionary XML file 162,870 lines (110,165 tags) 6.05 Mbytes
  - Supplement XML files: comparative lexicon, comparative sentences, placenames, ancestral beings, missing entries
-

# XML sample

```
4945 </entry>
4946 <entry lbl="66" num="66" suf="">
4947   <lemma>
4948     <di>bununu pirna</di>
4949     <pos val=""/>
4950   </lemma>
4951   <note type="editor" subtype="morphology"><ortho>pununu pirna</ortho></note>
4952   <gloss>large itch</gloss>
4953   <note type="idiom">This is used in a figurative sense of a man who rejects his wife, in
4954     order to look around for another [woman], and after that for [still] another. The idea
4955     is thereby expressed: 'He is a man driven by sensual appetite for one girl or woman
4956     after another'. Cf. <di>baka pirna</di>.</note>
4957   <note type="mythology">the term <di>bununu parana</di> relates to the <di>Mura</di><mura
4958     value="female">Warilani</mura>
4959     <fn author="Scherer" num="fn1"><note>In all probability this name should read <mura value="female"
4960       >Wariliwulani</mura>.</note></fn> an old woman who once had many sores and rashes
4961     (<di>tapa</di>). While these were [in the process of] healing,
4962     they [still] caused her considerable itching. So she used her hands to soothe this
4963     irritation of the skin. That is how the word is said to have originated.</note>
.....
```

---

# Strengths

- Huge number of entries, sub-entries, and examples gives rich information about possible semantics of Diyari lexicon, especially collocations and contexts
  - 610 idioms represented, many not present in modern sources
  - Rich encyclopedic information on culture and society (e.g. 915 ethnographic notes with 18 sub-types: 39 artefacts, 52 death, 30 kinship, 22 ceremony)
  - Basic information on mythology for 461 entries (cf. other Reuther volumes)
  - Comparative wordlist (157 entries), comparative sentences (10)
-

---

# Editorial interventions

- Typos corrected in Diyari, English, German
  - Where forms are known from modern sources they are added in the current orthography (for 2,486 entries)
  - Translator's English language errors corrected
  - Removed 'a' and 'the' in glosses, as per lexicographic practice
  - Scopal ambiguities resolved (e.g. multiple modifiers of head noun), e.g. 'clear, unobscured view' → 'clear view, unobscured view'
-

# Editing

- Contextual information (e.g. selection) moved from “gloss” to “context note”, e.g. ‘to duck from the boomerangs, (when they come flying through the air)’ → `<gloss>to duck from boomerangs</gloss><note type="context">when they come flying through the air.</note>`
- All/any of the above, e.g. ‘to eat, chew (e.g. tobacco) for someone else’ → `<gloss>to eat for someone else, to chew for someone else</gloss><note type="context">e.g. tobacco.</note>`; e.g. ‘to look well after (e.g. widows, children)’ → `<gloss>to look after well</gloss> <note type="context">e.g. widows, children.</note>`



# Editing

- glosses of idioms are clearly divided into literal gloss and idiomatic meaning, e.g. '[lit: to insult the water], i.e. not to take any fish out of forbidden waters' → 'to insult water'  
*Idiom*: not to take fish out of forbidden waters;
- Footnotes appear at their relevant location in the text (not at bottom of page) and are displayed as popups triggered by mouse over;
- Other editorial notes, e.g. essential clarifications, corrections of translator's comments in footnotes. Additions by the editor appear preceded by an asterisk

# Editing – sensitive expressions

- Reuther's dictionary contains expressions and opinions that were common among missionaries and other non-Indigenous people in the 19th and early 20th century which Aboriginal people and others may now find offensive.
- Examples are “witchdoctor”, “heathen”, “pagan”, and “native”. There are 386 instances of such potentially offensive expressions.
- In the XML file these are tagged and a substitute provided  
e.g.1 – like the cap on a <substitute val="">native</substitute>  
man's head  
e.g.2 – in conformity to <substitute val="the Law">pagan [or  
tribal] law</substitute>

---

# User interfaces

1. **Specialist edition** for researchers and advanced learners -- presents all material in Scherer translation with editorial amendments. It is formatted by a CSS that converts XML tags to display as format (colours, bold, italics, indentation, alignment, pop ups) in a browser
  2. Limited search capabilities within the browser, minimal use of hypertext
-

# Interface1

10

## **bakina vi**

\*Spelling: **paki-rna**

*to break open; to open up of its own natural force or instinct; to burst open; to crack; to burst asunder*

*Context:* Used in Diari of clouds, when they send forth rain in torrents, or when they disperse. Examples in Diari follow.

[1]

**tapa bakina warai**

*the sore has opened up; the wound has burst*

[page 11] [Vol. I, p. 9]

[2]

**nguramarali bakila wapaia**

*the rosy-fingered morning has dawned*

[3]

**nauja marda bakina warai**

*the stone has cracked*

*Context:* from the heat.

[4]

**nauja turu bakina warai**

*the fire has burst into flame*

*Context:* from wood laid on the coals

---

# Supplements (e.g. missing entries)

- Presented as per the Specialist edition but with lots of additional materials added by the editor, including references to additional sources, cognates in related languages, scientific identifications, hypertext links to Wikipedia entries, images, hypertext links back to the main dictionary
  - Sorted in regular alphabetical order – looks more like a “proper” dictionary
  - Balance between “faithfulness” and adding value
-

# Supplement

35

## **bakubaku** *n*

*\*Spelling: unknown*

*type of animal*

*\*Editorial Note: this is probably **pakupaku** 'crested bellbird'. See also Hercus (2014: 215) for the cognate term in Arabana-Wangkangurru.*

*\*Scientific name: Oreoica gutturalis*

*\*Scientific reference: [https://en.wikipedia.org/wiki/Crested\\_bellbird](https://en.wikipedia.org/wiki/Crested_bellbird)*

*[Wikipedia link \(opens new tab\)](https://en.wikipedia.org/wiki/Crested_bellbird)*



*\*Editorial Note: this item only appears as a sub-entry of **thidna** *foot**

*\*See: No. 3270-76*



---

# User interfaces

2. **User-friendly edition** for learners, community members, interested others -- presents all material in Scherer translation with editorial amendments, but without footnotes, page numbers, and with substitutions of sensitive vocabulary.
  
  3. Three interaction methods:
    - ❑ by letter groups that show Diyari vocabulary and English gloss only. Users can click to open up full entry display for items that interest them
-

# Letter group display for “b”

Reuther Dictionary

[Home](#)

[Dictionary](#)

[Notes](#)

[More](#)

Search

[a](#)

[b](#)

[d](#)

[g](#)

[j](#)

[k](#)

[m](#)

[n](#)

[ng](#)

[nj](#)

[p](#)

[t](#)

[tj](#)

[u](#)

[w](#)

**baka** n. *type, species; style, manner; nature, habit of a person or thing*



**bakajerrujerru** n. *hale and hearty nature*



**bakakaritjina** n. *transformed nature*



**bakana** conj. *too; also; and also; not only... but also*



**bakanamata** conj. *report; admission; confession; disclosure*



**bakanata** conj. *too; also; and also; not only... but also*





# Expand selected entry


Reuther Dictionary Home Dictionary Notes More

  
Search

a b d g j k m n ng nj p t tj u w

**baka** n. *type, species; style, manner; nature, habit of a person or thing* 

**bakajerrujerru** n. *hale and hearty nature* 

**bakakaritjina** n. *transformed nature* 

*Morphology:* paka kartyi-rna

*Mythology:* This word owes its origin to the **muramura Matjamarpina**, who had fat arms and legs, and was therefore **bakapilki**. One and the same expression is used in all dialects.

**jakaiai nauja karari bakakaritjina warai!**

*well, I never! he is now a transformed (almost disguised) type of man, (possibly because he has shaved off his beard)*

**bakana** conj. *too; also; and also; not only... but also* 

---

# For users

2. other interaction methods:
    - ❑ by search over Diyari or English expressions. The search box presents pop-up lists of available terms in forms or glosses and returns all entries with the search term – search begins with first typed letter and narrows as user types
    - ❑ by categorized Notes
-

# Search “do”

Reuther Dictionary Home Dictionary Notes More

Welcome to the home page for the Reuther Diyari Dictionary.

This is the online edited version of Rev. Philipp Scherer's (1981) English translation of the *German Dictionary*. It was created by Peter K. Austin, Edward Garrett, and David

Diyari (also spelled Diari or Dieri) is an Australian Aboriginal language spoken in the north-west of South Australia (see <https://www.diyari.org>).

To view the dictionary click **Dictionary** in the navigation bar and you will see a list of words beginning with that letter. Click on a different letter to see words beginning with that letter. If you scroll over a word, a pop-up window will open up and display its full dictionary information.

To look for English or Diyari words in the dictionary use the **Search box**. Type the word you are interested in and you will see a drop-down list of words in the Dictionary that begin with that letter. Click on the word you want and click Search. You will be presented by all occurrences of the search word in the Dictionary. You can click on the highlighted links in each search result to go to the relevant dictionary entry.

A classified list of all the notes in the Dictionary can be found under the **Notes** tab.

dog  
doubter  
dokupirra  
camp dog  
doku belu  
doku manka  
doku maika  
doku karia  
lidna doku  
marda doku  
doku palara  
doku parina  
doku kekilja  
doku palparu  
worker, doer  
doku tjinpiti

do



Search

Reuther's 1908 *Diari-*

ist of South Australia (see

beginning with the letter 'a'.  
a particular word it will

s of the word you are  
letters. Choose the one  
; and examples. You can  
he examples of their use.

# Search “dok”

Reuther Dictionary Home Dictionary Notes More

Welcome to the home page for the Reuther Diyari Dictionary.

This is the online edited version of Rev. Philipp Scherer's (1981) English translation of the *German Dictionary*. It was created by Peter K. Austin, Edward Garrett, and D

Diyari (also spelled Diari or Dieri) is an Australian Aboriginal language spoken in the north-western part of South Australia (see <https://www.diyari.org>).

To view the dictionary click **Dictionary** in the navigation bar and you will see a list of words beginning with that letter. Click on a different letter to see words beginning with that letter. If you scroll down the list will open up and display its full dictionary information.

To look for English or Diyari words in the dictionary use the **Search box**. Type in the word you are interested in and you will see a drop-down list of words in the Dictionary that begin with the letters you want and click Search. You will be presented by all occurrences of the word in the dictionary. You can click on the highlighted links in each search result to go to the relevant dictionary entry.

A classified list of all the notes in the Dictionary can be found under the **Notes** link in the navigation bar.

dokupirra  
doku balu  
doku manka  
doku malka  
doku karla  
doku palara  
doku parina  
fidna doku  
marda doku  
doku palparu  
doku kekijja  
doku kalikali  
doku tjinjiri  
doku njinjaru  
doku karitjina  
dokupirapirra

dok



Search

Reuther's 1908 *Diari-*

Dictionary of South Australia (see

the introduction beginning with the letter 'a'.  
If you click on a particular word it will

display all occurrences of the word you are  
interested in. Choose the one you want  
; and examples. You can click on the  
examples of their use.



# Search “dog”

dog  
camp dog  
wild dog; dingoo  
name of dog of muramura

Welcome to the home page for the Reuther Diyari Dictionary.

This is the online edited version of Rev. Philipp Scherer's (1981) English translation of Rev. J.G. Reuther's 1900 *German Dictionary*. It was created by Peter K. Austin, Edward Garrett, and David Nathan.

Diyari (also spelled Diari or Dieri) is an Australian Aboriginal language spoken in the far northeast of South Australia (see <https://www.diyari.org>).

To view the dictionary click **Dictionary** in the navigation bar and you will see a list of entries beginning with the letter 'a'. Click on a different letter to see words beginning with that letter. If you scroll down and click on a particular word it will open up and display its full dictionary information.

To look for English or Diyari words in the dictionary use the **Search box**. Type the first few letters of the word you are interested in and you will see a drop-down list of words in the Dictionary that begin with those letters. Choose the one you want and click Search. You will be presented by all occurrences of the search word in entries and examples. You can click on the highlighted links in each search result to go to the relevant dictionary entries or to the examples of their use.

A classified list of all the notes in the Dictionary can be found under the **Notes** tab.

# Notes list

## All notes

Notes have been classified by type and sometimes also subtype. Use the links below to browse notes of a specific type. Counts are shown in parentheses.

### **Comparative (138)**

### **Context (4120)**

### **Editor (4940)**

→ alternative (15)

→ grammar (3)

→ morphology (4260)

→ semclass (358)

→ xref (110)

---

# Challenges of Reuther's dictionary

- Reuther was a rather poor linguist, a weak lexicographer, and obsessed with arcane knowledge (ancestral beings, mythology, placenames) while uninterested in many aspects of the mundane world (the 'lone scholar')
  - He received little training in languages, other than Greek and Latin, and for Diyari grammar borrowed heavily from previous missionaries, such as Flierl (Stockigt 2016)
  - Reuther's missionary orthography over-differentiates vowels and under-differentiates consonants, resulting in distinct words being spelled the same, or the same sound spelled differently
-

# Phonology

---

*Reuther*

*Austin*

*Gloss*

**ngatata**

ngardarda

‘maternal grandfather’

**ngatata**

ngathata

‘younger brother’

**terti**

thati

‘middle’

**terti**

thardi

‘thirsty’

**tala**

darla

‘skin’

**tala**

tharla

‘name’

**kalu**

kalhu

‘liver’

**kalu**

karlu

‘testicles’

---

# Phonology

---

<b>kati</b>	kathi	'clothing'
<b>kadi</b>	karti	'raw, uncooked'
<b>kadi</b>	kardi	'brother-in-law'

---

<b>ngura</b>	ngura	'camp'
<b>ngura</b>	ngurra	'continuous'

---

<b>baru</b>	paru	'yellow'
<b>paru</b>	parru	'fish'

---

<b>waka</b>	waka	'small'
<b>wokara</b>	wakarra	'neck'
<b>woma</b>	wama	'carpet snake'
<b>wapana</b>	waparna	'to walk'

---

# Morphology – derived forms

---

<b>buljubulju</b>	pulyupulyu	‘annoyed, sullen’
<b>buljubuljujeli</b>	pulyupulyu-yali	‘angry (transitive)’
<b>buljubuljurina</b>	pulyupulyu-ri-rna	‘to become angry’
<b>buljubuju ngankijirbamalina</b>	pulyupulyu-nganka-iyirpa-mali-rna	‘to complain against one another’
<b>buljubuljurilkijiribamalina</b>	pulyupulyu-ri-lka-iyirpa-mali-rna	‘to annoy each other’
<b>buljubuljurinietja</b>	pulyupulyu-ri-rna-yitya	‘surly type’
<b>buljubulju wapana</b>	pulyupulyu wapa-rna	‘to walk along sullenly’
<b>buljubulju kurana</b>	pulyupulyu kurra-rna	‘to devise dissention’

---



# Lexicographic issues

- Entries are sometimes partially duplicated, e.g.
  - **dapana** ‘to drink; to suck, to suck up; to kiss; to chew; to eat (of grounded seed); to belch or burp; to wet or moisten; to pour, to swallow’
  - **tapana** ‘to drink’, but its sub-entries contain glosses ‘to slurp, to suck, to absorb (water), chew, lick’
- Sub-entries are often randomly listed in an apparent stream of consciousness (by Reuther or his teachers)
- Some common terms have dozens of sub-entries, with tenuous semantics, e.g.
  - **tidna** ‘foot’ has 223 sub-entries, many of which are names for animals that have tracks
  - **tandra** ‘fruit’ has 127 sub-entries, most of which are names for unidentified plants

# Lexicographic challenges

- some relatively common words are mistranslated, e.g.
  - **kaku** ‘sister (brother speaking)’ → ‘older sister’
  - **ngatata** ‘younger brother’ → ‘younger sibling’
  - **kami** ‘paternal grandmother’ → ‘father’s mother, father's mother’s brother’
- over 300 items do not have a headword entry but appear only as a sub-entry, often under a semantically loosely connected headword, or in examples, e.g. **tindritindri** ‘willy wagtail’ appears under **tidna** ‘foot’ only (as an animal that has tracks), **karku** ‘red ochre’ appears in 33 sub-entries and 10 examples but has no entry itself (c.f. **bukatu** ‘pink ochre’)
- Entries Supplement for these – many terms refer to flora and fauna, which Reuther apparently had no interest in but for some we can identify from other sources, some contemporary to him, e.g. Gason (1886)
- Significant sense, reference, and cultural information can be scattered across entries in notes or examples

---

# Proposal

- The main Reuther-Scherer dictionary needs a **stand-off XML index** that also incorporates material from all other legacy and modern sources (something like what we have done for the missing items Supplement)
  - It should be properly sorted and structured lexicographically to create a stronger Diyari-English bilingual encyclopaedic dictionary that builds on Reuther's strengths while filling the gaps and correcting his errors (especially in phonology and morphology)
  - We are working on this currently using Toolbox and Lexique Pro to generate both printable and web accessible versions
-

\x Lexeme

ngaRu

\a Alternate

ngaRu

\c Citation form

0527

\var Variant

Di

\xnum Lexical number

\g Language

ʔ

\vcl Verb Class

voice

\ps Part of Speech

\ge Gloss in English

echo

\ge Gloss in English

\de Definition

voice, distinctive sound, echo

\eth Ethnographic

\sci Scientific name

\scisrc Scientific source

\nt Note

Trefry (1984: 181) |

\sd Semantic domain

[Vocalisation\\_thought](#)

\syn Synonym

\ant Antonym

\cf Cross-reference

**karta, kaldra, ngayarla, kunngaRa**

\excf External cross-reference

**ngaRu**

\gcf Language cross-reference

YY

\reu\_num

2049

\reu\_lx

ngaru

\reu\_ge

\bv Berndt and Vogelsang

\howitt Howitt form

\gas\_lx Gason form entry

\rec Recorder

PKA, DT

\sp Speaker

\x\_ref Example reference

[Di-g456](#)

\xv Example vernacular

**ngathu nhinhayari ngaRu ngaRarna warayi**

\xe Example English

I heard a voice like his

\date Date

04/Aug/2023

(a) 'to drink'

*ngapa thaparna* 'to drink water' [R.229-ex1, R.3148-ex1]

*kumarrri thaparna* 'to drink blood', *Ethn.* When he is first being made a man, the *wiyarru* opens his mouth widely and drinks the blood offered to him by his initiators. This also used to refer to drinking the blood that has been washed off a spear that a man has been killed with; young men are required to do this, in order to become fearless. Reuther also has the meaning 'to suck a wound' [R.3148-ex 9]

*kipaya thaparna* 'to drink urine', *Ethn.* men drink their own urine in order to end a friendship when planning to kill a friend [R.3148-ex22]

*piarrurna thaparna* 'to kneel down to drink' [R.3148-ex18]

*thapatha thikarna* 'to come back to drink' [R.3148-ex21]

(b) 'to suck'

*ngama thaparna* 'to suck the breast' [R.3148-ex6]

*paya kapi thaparna* 'to suck out bird's eggs' [R.3148-ex7]

*thurintyi thaparna* 'to suck marrow out of a bone' [R.3148-ex8]

*muanya thaparna* 'to suck a wound', *Ethn.* the *kunkri* sucks puss and matter from a wound [R.3148-ex3]

*mularru thaparna* 'to suck on caterpillars', *Ethn.* this is done so they become big and fat [R.3148-ex19]

(c) 'to eat, slurp up, or chew on soft or semi-liquid food or fruit'

*pawa thaparna* 'to slurp up ground seed' [R.229-ex6, R.3148-ex2]

*kilthi thaparna* 'to eat fat or stew' [R.229-ex4, R.3124-ex 5]

*danyu thaparna* 'to eat *danyu* fruit'. *Note.* for soft fruit *thaparna* 'drink' rather than *thayirna* 'eat' is used. Other fruits are *kudnampira*, *mpiampia*, *nharimayi*, *ngalyaru*, and *ngurathikiri* [R.3124-ex11]

*palyangari thaparna* 'to chew on the gum of the *palyangari* tree' [R.229-ex9, R.3148-ex 12]

*karna thaparna* 'to eat people', *Ethn.* the members of a *pinya* revenge expedition eat the raw liver of their victim. [R.229-ex2, R.3148-ex14]

(d) 'to lick'

*miralu thaparna* 'to lick a coolamon', *Ethn.* when a man has no tobacco, he licks the dish that he last prepared tobacco in to ensure that some may soon become available. [R-ex 23]

(e) 'to kiss, touch with the mouth'

*marna thaparna* 'to kiss on the mouth'

*parru thaparna* 'to kiss a fish', *Ethn.* If no fish land in the net, a man goes down into the water, whistles into a hollow bone, and sings his *mura* song. The first fish to be caught is then kissed while the man has bread (made of seed) in his mouth, and allowed to swim again. This is expected to entice other fish to enter the net. [R.3148-ex20]

*kira thaparna* 'to kiss a boomerang', *Ethn.* this is done before throwing to ensure that the boomerang hits its target. It is also done repeatedly when a boomerang is being made to see if it is straight. [R.229-ex25, R.3148-ex13]

---

# Conclusions

- creating and analysing corpora can be very rewarding and enable various exciting kinds of linguistic and cultural research to be done
  - however, working with corpora, especially legacy materials, involves dealing with often **complex issues** about the form, content, context, and use of materials and analyses arising from them
  - maximising opportunities to use a corpus requires thinking about data entities, data types and relationships, and being **explicit** about them in the project design and application (e.g. in database design or XML tagging)
  - very important role for **metadata** and **meta-documentation**
  - by creating good meta-documentation now we can reduce legacy data problems for future researchers
-

---

# Conclusions

- there are many **opportunities** for researchers to add substantial value to legacy corpus materials, and create **secondary** corpora, especially if they are able to work with other historical sources and/or contemporary knowledge holders to elucidate them and the context surrounding their creation, analysis and current status
  - careful work with legacy corpora can also be very **rewarding** for researchers and communities, especially for unique documents on languages/varieties or areas of knowledge that are no longer available, and that can serve as **important sources** for language support and revitalisation
  - Thank you for your attention
-



---

# References

- Austin, Peter. 1981. *A grammar of Diyari, South Australia*. Cambridge: Cambridge University Press.
- Austin, Peter K. 2014. And still they speak Diyari: the life history of an endangered language. *Ethnorema* 10, 1-17.
- Austin, Peter K. 2013. Language documentation and meta-documentation. In Mari Jones & Sarah Ogilvie (eds.) *Keeping Languages Alive: Documentation, Pedagogy and Revitalization*, 3-15. Cambridge: Cambridge University Press.
- Austin, Peter K. 2021. *A Grammar of Diyari, South Australia*. 2nd edition, version 2.10. Manuscript, SOAS, University of London.
- Austin, Peter K. 2023. Making 2,180 pages more useful: the Diyari dictionary of Rev. J. G. Reuther. In Eda Dehermi & Christopher Moseley (eds.) *Endangered Languages in the 21st Century*, 241-257. London: Routledge.
- Christen, Kim. 2018. Relationships, not records: Digital Heritage and the Ethics of Sharing Indigenous Knowledge Online. In Jenetry Sayers. (ed.) *Routledge Companion to Media Studies and Digital Humanities*, 403-412. London: Routledge.
- Dobrin, Lise & Saul Schwartz. 2021. The social lives of linguistic legacy materials. *Language Documentation and Description* 21, 1-36.
-

---

# References

- Hercus, Luise. 2017. Looking at some details of Reuther's work. In Nicolas Peterson and Anna Kenny (eds.) *German Ethnography in Australia*, 115–135. Canberra: ANU Press.
- Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36, 161–195.
- Himmelman, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation and Conservation* 6, 187-207.
- Nathan, David & Peter K. Austin. 2004. Reconceiving metadata: language documentation through thick and thin. *Language Documentation and Description* 2, 179-187.
- Nathan, David, Susannah Rayner & Stuart Brown (eds.) 2009. *William Dawes : notebooks on the Aboriginal language of Sydney: a facsimile version of the notebooks from 1790-1791 on the Sydney language written by William Dawes and others*. London: SOAS. (see also [www.williamdawes.org](http://www.williamdawes.org))
-

---

# References

- Reuther. J. G. 1901. A Diari dictionary. Manuscript translated by Phillip A. Scherer, held at the Australian Institute of Aboriginal and Torres Strait Islander Studies..
- Reuther, Johann G. 1981. *The Diari vols 1-13*. Translated by Philipp A. Scherer. Vol. 5 translated by T. Schwarzschild and L.A. Hercus. AIAS Microfiche No.2, Canberra: Australian Institute of Aboriginal Studies.
- Reuther. J. G. & Carl Strehlow. 1897. *Testamenta marra. Jesuni Christuni ng ngantjani jaura ninaia karitjimalkana wonti Dieri jaurani*. Adelaide: G. Auricht.
- Stockigt, Clara. 2016. Pama-Nyungan morphosyntax: lineages of early description. PhD thesis. Adelaide University.
- Warner, N., Q. Luna, and L. Butler. 2007. Ethics and revitalization of dormant languages: The Mutsun language. *Language Documentation & Conservation* 1(1), 58-76.
- Warner, N., Q. Luna, L. Butler & H. van Volkinburg. 2009. Revitalization in a scattered language community: Problems and methods from the perspective of Mutsun language revitalization. *International Journal of the Sociology of Language* 198, 135-148.
-