LDSS2025 seminar, 17 July 2025

Challenges and opportunities in dealing with legacy materials

Peter K. Austin
Department of Linguistics
SOAS, University of London

www.peterkaustin.com

General issues for discussion

1. What do we mean by ‘legacy materials’?

2. Some examples

3. Challenges of working with legacy materials

4. Opportunities to add value to a legacy corpus – some case studies

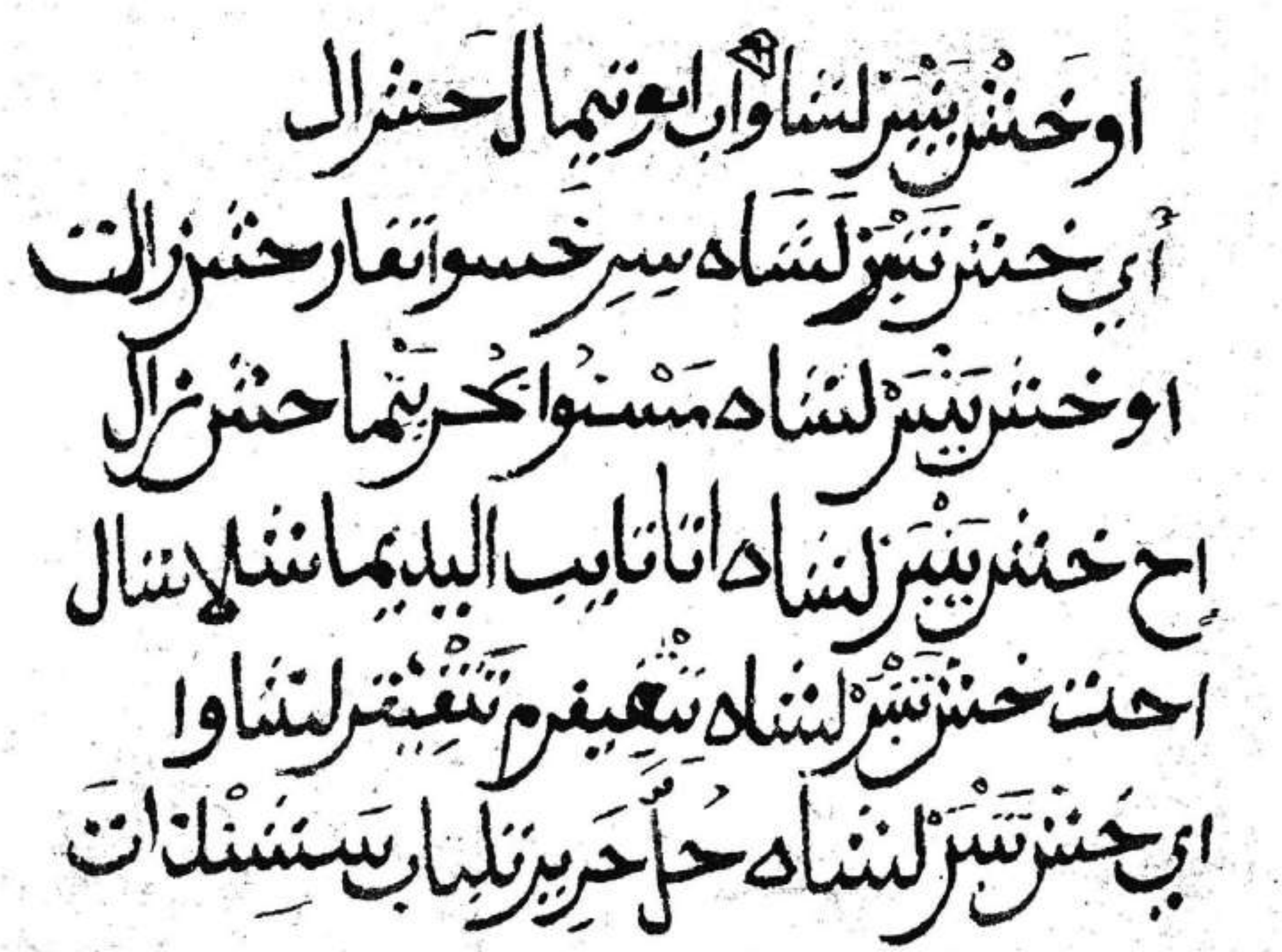
5. Outcomes of legacy materials research

Legacy materials

I use this term to refer to any language/dialect-related material that was **not** created as part of a current project:

- **Types**; audio or video tapes, written or typed manuscripts
- **Sources**: travellers, colonial officials, missionaries, community members (authors, poets, story tellers, biographers)
- **Genres**: report, academic paper, diary, poem, novel, description (dictionary, grammar), fieldnotes
- **Metadata**: typically minimal so may require historical research and/or guesses/assumptions from the form and/or content of the materials

Legacy materials examples



او خنش بنش لسناو اب اعو تيما لحنزال

أيِ خنش نبش لساه سِر خسو اتقار حنز الت

او خنش بنش لساه مسنو اكحر بنما حنش زال

إح خنش بنش لساه اتاتابب البدبما سنالا اتنال

احت خنش نبش لساه نعيفر نتعبثر لتساوا

ايَ خنش نبش لساه حل حربر تلياب سنسنلات

Legacy materials example 1

- From Prof Giorgio Banti
- Old Harari, Ethio-Semitic language, written in Ajami Arabic, used for religious texts of different genres since the 15th century
- Several manuscripts in two libraries in Rome, worked on them in Rome and Harar (Ethiopia)
- Fieldwork on performances: audio recording with solo, drums and chorus during a *zikri* ceremony in Harar about a decade ago, pictures of the same lines in an Ajami manuscript, the Latin transcription and the translation
- Preparations for its publication are underway

[illegible]	[illegible]	[illegible]			
[illegible]	[illegible]	[illegible]			
[illegible]	[illegible]	[illegible]			
	[illegible]				
conférer	[illegible]	[illegible]	[illegible]	[illegible]	
conforme	[illegible]	[illegible]			
congé	[illegible]	[illegible]	[illegible]		
congédier	[illegible]	[illegible]	[illegible]		
conjurer	[illegible]	[illegible]	[illegible]		
connaissance	[illegible]	[illegible]		[illegible]	
connaître	[illegible]	[illegible]	[illegible]	[illegible]	
congrégation	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]		[illegible]	
[illegible]	[illegible] [illegible]	[illegible]		[illegible]	
conseil	[illegible]	[illegible]	[illegible]	[illegible]	
conseiller	[illegible]	[illegible]	[illegible]	[illegible]	
consentir	[illegible]	[illegible]	[illegible]	[illegible]	
consentement	[illegible]	[illegible]	[illegible]	[illegible]	
conséquence	ፍሬ	[illegible]	[illegible]	[illegible]	
conservation	[illegible]	[illegible]	[illegible]	[illegible]	
conserver	[illegible]	[illegible]	[illegible]	[illegible]	
considération	[illegible]	[illegible]	[illegible]	[illegible]	
considérer	[illegible]	[illegible]			
consoler	[illegible]	[illegible]	[illegible]	[illegible]	
consolation	[illegible]	[illegible]	[illegible]	[illegible]	
consolider	[illegible]	[illegible]	[illegible]	[illegible]	
consommer	[illegible]	[illegible]		[illegible]	
	[illegible]			[illegible]	
	[illegible]			[illegible]	
consomption	[illegible]	[illegible]	[illegible]	[illegible]	
être consumé	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]			
construire	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]				
consul	[illegible]	[illegible]	[illegible]	[illegible]	
consulter	[illegible] [illegible]	[illegible]	[illegible] [illegible]	[illegible] [illegible]	
consumer	[illegible]	[illegible]	[illegible]	[illegible]	
contact	[illegible]				
conte	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]	[illegible]		[illegible]	
contenir	[illegible]	[illegible]	[illegible]	[illegible]	
être content	[illegible]	[illegible]	[illegible]	[illegible]	
contre: [illegible]	[illegible]	[illegible]		[illegible]	
continuer	[illegible] [illegible]	[illegible]	[illegible]	[illegible]	
contester	[illegible]	[illegible]	[illegible]	[illegible]	
contempler	[illegible] ~~[illegible]~~	[illegible] ~~[illegible]~~		~~[illegible]~~	
contract	[illegible]	[illegible]	[illegible]		
contracter	[illegible] [illegible]	[illegible]	[illegible]	[illegible]	
contrainte	[illegible]	[illegible]		[illegible]	
contraindre [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
contrée	[illegible]	[illegible]	[illegible]	[illegible]	
	[illegible]			[illegible]	
contrefaire: [illegible]	[illegible]	[illegible]	[illegible]	[illegible]	
contrevenir	[illegible]	[illegible]		[illegible]	
contre-poids	[illegible]		[illegible]	[illegible]	
contribution	[illegible]	[illegible]	[illegible]	[illegible]	
contrition	[illegible]	[illegible]	[illegible]	[illegible]	
controuver	[illegible]	[illegible]	[illegible]	[illegible]	
contusion	[illegible]	[illegible]	[illegible]	[illegible]	
[illegible]	[illegible]	[illegible]			
[illegible]	[illegible]	[illegible]			
[illegible]	[illegible]	[illegible]			

Legacy materials example 2

- From Prof Moreno Vergari
- Handwritten 1910 wordlist > 4,000 words, in five languages (French, Tigrinya, Saho, Amharic, and Ge’ez). The three Semitic languages and the Saho are written in Ethiopic script
- Author: Joseph Baeteman, Catholic Lazarist missionary, worked among Irob of northern Ethiopia, speaking a southern variant of Saho
- Found and photographed this manuscript in 2020 in a monastery in Tigray. At the end, gave the monastery leaders a CD with all the photos. The originals are at great risk of deterioration, and also from possible looting in the library by soldiers during the constant fighting there
- Currently using them for preparing an article (“The Saho of Eritrea and Ethiopia: the history of the studies”) that will be published in the book *The Routledge Handbook of the history of descriptive linguistics.*

Vt. depois de க. escreuem ங. este ங. ase de por antes de க. Depois de ச. escreuem ஞ. este ஞ ase de por antes de ச. &cª Vt.

ங்க.	Vt.	முங்கில்.	munguil.	Bambu.
ஞ்ச.	Vt.	தஞ்சம்.	Tanjam.	Emparo.
ண்ட.	Vt.	பண்டம்.	Pandam.	Fato.
ந்த.	Vt.	தந்தம்.	Tandam.	Marfim.
ம்ப.	Vt.	உடம்பு.	Vdambu.	Corpo.
ன்ற.	Vt.	பன்றி.	Pandi.	Porco.

Tanchjam

Nota.

Quando algum nome acaba em. ர. ல. ள. ன= se lhe presede a breue pruncia sse como se ari el alli an idest como a intermetendo: e. Vt.

அவன். auan.
அவர். auar.
அவள். aual.
வாராமல். varamal.

No escreuer tamul quando hum nome acaba em vogal e se segue. க ச ப த. ordinariamente escreuem estes consoantes dobrados. Vt.

Quando hum nome acaba em ல. simplex e se segue ப hase de por ப்ப. dobrado Vt.

Quando hum nome acaba em vogal e se segue otro por vogal ordinariamente se fas sinalefa

Quando hum nome acaba em vogal e se segue vogal se naõ quisermos fazer sinalefa podemos escreuer com ய. யா. யி. யீ. ett. ou com வ. வா. வி. வீ. &cª. Vt

. Plura his tibi suggeret vsus.

1665

பிலீபபிவல்தெயுசு

Legacy materials example 3

- From Prof Cristina Muru
- handwritten manuscript (Cod. Orient. 283) found in the Hamburg State and University Library in Germany in 2016
- it belonged to a Dutch Calvinist clergy Philippus Baldaeus (1632-1672) and contains a Tamil grammar (inspired) by the Jesuit missionary at the mission of Cardiva, Gaspar De Aguilar, dismissed from the Company of Jesus in 1645
- demonstrates how Protestants relied on the Jesuits’ materials for implementing their linguistic and religious activity in India
- part of ongoing work on the history of European research in southern India

Working with legacy materials

1. **Challenges**: practical, technical, ethical, and political issues, and many questions which it may be difficult or even impossible to answer
2. **Opportunities:** to **add value** to legacy materials using documentary linguistics methods
3. Distinguish between adding **structure** (categories, entities, relationships), adding **content**, and adding **format**
4. Explicit and well-structured data (e.g. stored in a database, or marked up in extensible markup language (XML)) can have **format added** computationally, and also lends itself to **repurposing** for other uses and/or other users than the immediately intended audience (and so can be multifunctional, as language documentation proposes)

Reference: Austin, Peter K. 2017. Language documentation and legacy text materials. *Asian and African Languages and Linguistics* 11, 23-44

Working with legacy contexts

Essential to explore the socio-cultural and historical context of the documents and their creation:

1. biography of the author(s), prior language knowledge and/or study and/or exposure, their teachers/mentors/correspondents, duration of work on the language and career stage, what funding, what goals, previous studies of the language or the community that they could have had access to
2. explore aspects of historical period during which materials were created: kind and impact of contact between communities, including colonialists, and what descriptive categories and formats would have been known and might have influenced the author(s)

There may be ethical issues about informed consent (at the time of collection, by modern descendants), and access and use of legacy materials (“data sovereignty”)

Working with form

1. Form issues: handwriting, unfamiliar scripts, shorthand, author oddities and inconsistencies (paleographic, philological, and linguistic training necessary)
2. Orthography may over-differentiate or under-differentiate sounds (Austin 2023 on Reuther's Diyari)
3. Textual amendments (crossing out, additions), abbreviations, or other obscurities
4. Typesetting/scanning/OCR errors for publications – return to manuscripts if possible
5. Link transcription to images of the original documents so that readers may confirm the proposed analysis (cf. https://www.williamdawes.org/)

Working with form

1. **Diplomatic edition** aims to accurately reproduce all significant features in the original manuscript, including spelling and punctuation, abbreviations, deletions, insertions, and other alterations, e.g. https://www.utad.pt/cel/wp-content/uploads/sites/7/2022/10/CEL_Lingui%CC%81stica_25_revised.pdf
2. Text material may use typography or page layout to **implicitly** code information – we should represent this **explicitly** in our data model. Relies on socio-cultural and linguistic **literacy**

Working with content

1. cryptic glossing, or wrong glosses, because the author could not understand their language consultant’s accent or pronunciation, or because the semantics of the source language terms were misunderstood – Austin and Crowley (2005) LINK
2. Changes in understanding over time -- different parts of a collection of text material may show different spellings, translations, analyses etc.
3. Some content may be inappropriate to discuss/display to particular individuals or groups within a community (e.g. children, women)
4. Content can be dated, offensive or inappropriate by contemporary standards – Austin (2023) mentions “pagan”, “heathen”, “primitive”
5. Author personal comments that were never intended for publication

Working with analysis and context

1. Author may **record** a distinction (phonological, morphological, syntactic) that is not actually present in the language but may exist in the author’s 1st or other language, e.g. missionaries record “Vocative” case form in Diyari which is shouted form of any word with final vowel distortion
2. Author may **miss** a crucial distinction that does not exist in own or known languages, e.g. velar vs glottal/uvular stop, tone, Diyari missionaries ‘Modus Conditionalis’ shows switch-reference system (same-different subject)
3. Lack of metadata and meta-documentation can mean **unclarity** of:
    - **a.** **speaker**: age, geographical origin, social position, relationships to other contributors, languages learnt as 1st or 2nd language/dialect
    - **b.** **collector**: background, languages known, education history, prior language (and linguistic) study, research training and methodology, research methods and tools, familiar books and articles, teachers and interlocutors (mentors, colleagues), research goals, career trajectory, relationships with language consultants and community
4. Construction of knowledge is socially, culturally and historically embedded

Working with stakeholder issues

1. Need to identify who has an interest in the documents: author-side, community-side
2. Changes in community membership over time, who has authority?
3. Agreements about access and use, including publication (then, now): open, restricted, request, closed
4. Determining rights: intellectual property, copyright, moral (affecting reputation), inheritance of rights, orphan works
5. In analysis and publication clearly document who did what (e.g. in tags or data fields)
6. May require sensitive negotiations and extensive discussions with stakeholders, often over an extended time period, and may change later

Representational methods

There are two major ways that data can be represented:

a. **Relational** model – entities and relationships between them (1-to-1, 1-to-many, many-to-many). Typically encoded as a set of tables using a relational database software (cf. Austin 2025 on Wurm's Guwamu fieldnotes)
b. **Extensible Markup Language** (XML) model – hierarchical specification of entities and their content properties using pairs of tags, e.g. <entry><gloss> … </gloss></entry> (cf. Austin 2023 on Reuther's Diyari dictionary)

- Underlying digital representation in XML format can distinguish content structures, categories, information types, and layout structures
- Use Extensible Stylesheet Language Transformations (XSLT) to select or change the content of one XML document to another
- Use Cascading Style Sheets (CSS) to lay out the text visually on the page or screen

3148.

IV,80	tapana$ (v) = ‘to drink’
	1) ngapa tapana = ‘to drink water’
	2) paua tapana = ‘to slurp [or sip] seed-pulp’
	3) muntja tapana = ‘to suck on a patient’, i.e. to suck out the rubbish[2] at the seat [or source] of the trouble. The kunki does this.
	4) mitali ngapa tapana = ‘for the ground to absorb the water’
	5) gildi tapana = ‘to drink the fat’
	6) ngama tapana = ‘to suck [at] the breast’
	7) paja kapi tapana = ‘to suck out birds’ eggs’
	8) durintji tapana = ‘to suck the marrow out of a bone’

\fn1. Reuther: "wenn bei jemanden zum einen das andere kommt," – whatever that may mean.

\fn2. Reuther: "Unrat." P.A.S.

XML case study

Discussion: what is this?

What structures are present?

What format is present?

How do you know? What literacy skills are you using?

Case study

A page from Scherer’s translation of Reuther’s Diyari-German dictionary

Hierarchical Structure:

1. Page with Scherer number, Reuther manuscript, footnotes
2. Lexical entry, which contains forms, glosses (translations), part of speech, numbered subentries (forms, glosses), notes

Format:

1. Diyari words underlined
2. Glosses enclosed by = ‘…’
3. Notes begin i.e. …
4. German words in “…”

Case study

Let’s begin to specify XML

```
<dictionary title=“Reuther Diyari Dictionary”>
<page scherer_num=“1885”, reuther_num=“IV,80”>






</page>
</dictionary>
```

3148.

IV,80	tapana$ (v) = ‘to drink’
	1) ngapa tapana = ‘to drink water’
	2) paua tapana = ‘to slurp [or sip] seed-pulp’
	3) muntja tapana = ‘to suck on a patient’, i.e. to suck out the rubbish[2] at the seat [or source] of the trouble. The kunki does this.
	4) mitali ngapa tapana = ‘for the ground to absorb the water’
	5) gildi tapana = ‘to drink the fat’
	6) ngama tapana = ‘to suck [at] the breast’
	7) paja kapi tapana = ‘to suck out birds’ eggs’
	8) durintji tapana = ‘to suck the marrow out of a bone’

\fn1. Reuther: "wenn bei jemanden zum einen das andere kommt," – whatever that may mean.

\fn2. Reuther: "Unrat." P.A.S.

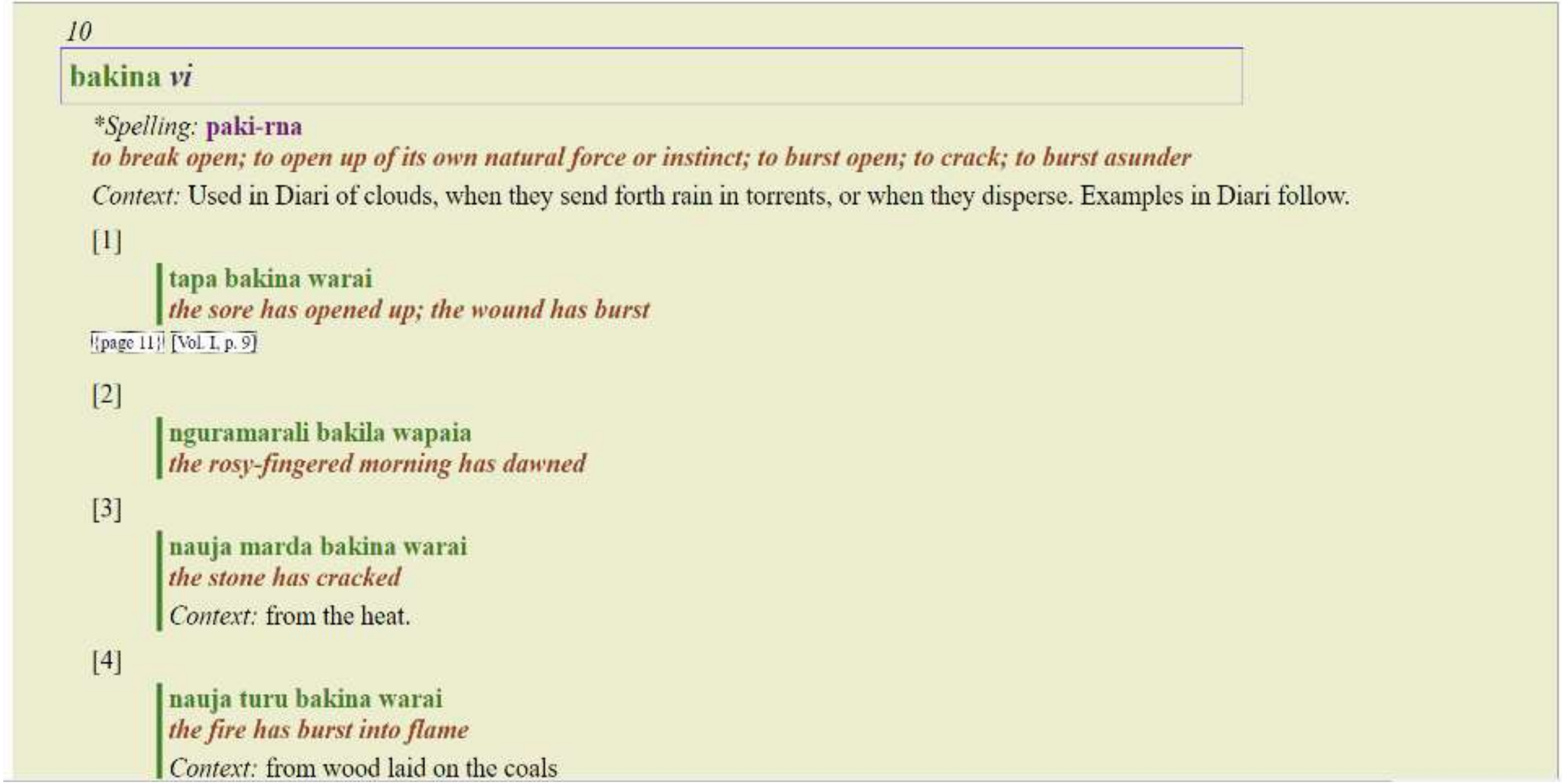
Case study

```
<dictionary title=“Reuther Diyari Dictionary”>
<page scherer_num=“1885, reuther_num=“IV,80”>
<entry num=“3148”>
<form lang=“DIF”>tapana</form>
<pos>v</pos>
<gloss>to drink</gloss>
<subentry num=“1”>
<form lang=“DIF”>ngapa tapana</form><gloss>to drink water</gloss>
</subentry>
<subentry num=“2”>
<form lang=“DIF”>paua tapana</form><gloss>to slurp [or sip] seed-pulp</gloss>
</subentry>
…
</subentry>
</entry>
</page>
</dictionary>
```

Case study outputs

1. **Specialist edition** that aims to be diplomatic edition including everything in the original manuscript plus author additions (modern spellings, corrections of errors)
2. **User-friendly edition** that aims at language learners and/or community members, contains less information and displays it in a different way emphasising ease of access
3. Specialist edition uses CSS to control formatting on a single large web page
4. User-friendly edition uses XSLT to select some information, then loaded into an online database and displayed with interactive capabilities, including predictive search on both the Diyari and English forms
5. Does not use a specialised program or app. Delivered as website that can be accessed on desktop or laptop computer, or on mobile device (tablet, phone) – see www.diyari.org, especially https://diyari.org/resources/reuther-dictionary/

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***

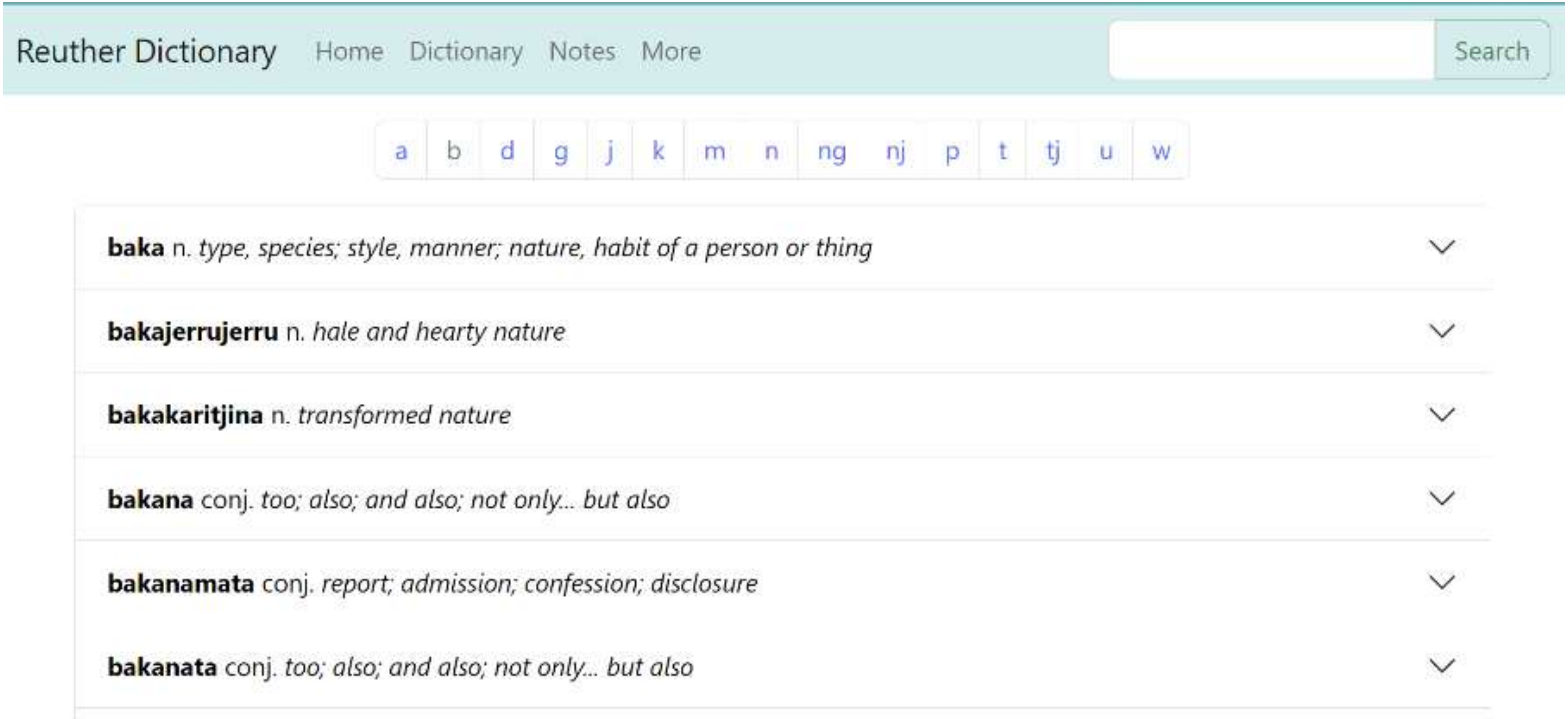
[3]

> **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

Letter group display



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

bakajerrujerru n. *hale and hearty nature*

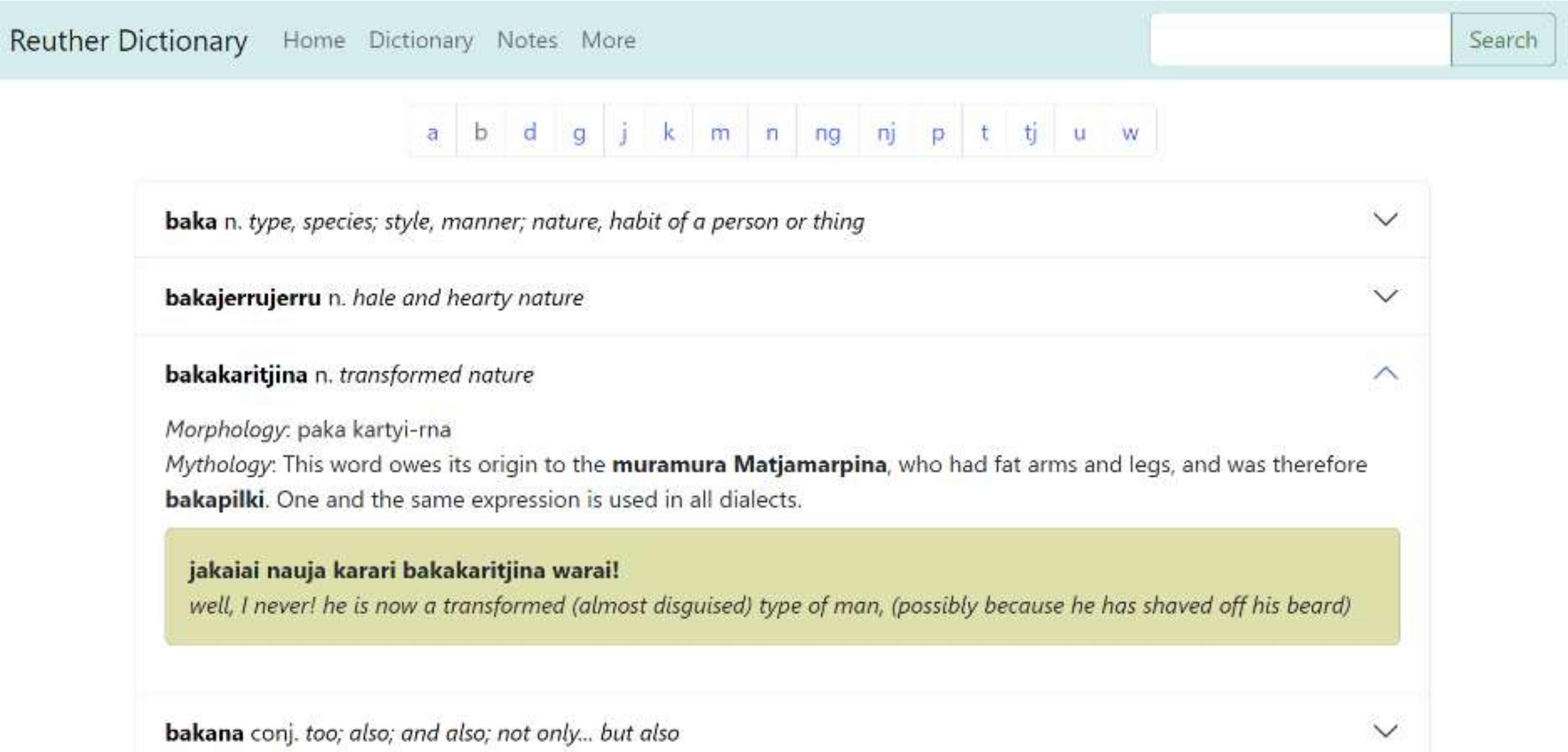
bakakaritjina n. *transformed nature*

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*

Expanded entry



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

bakajerrujerru n. *hale and hearty nature*

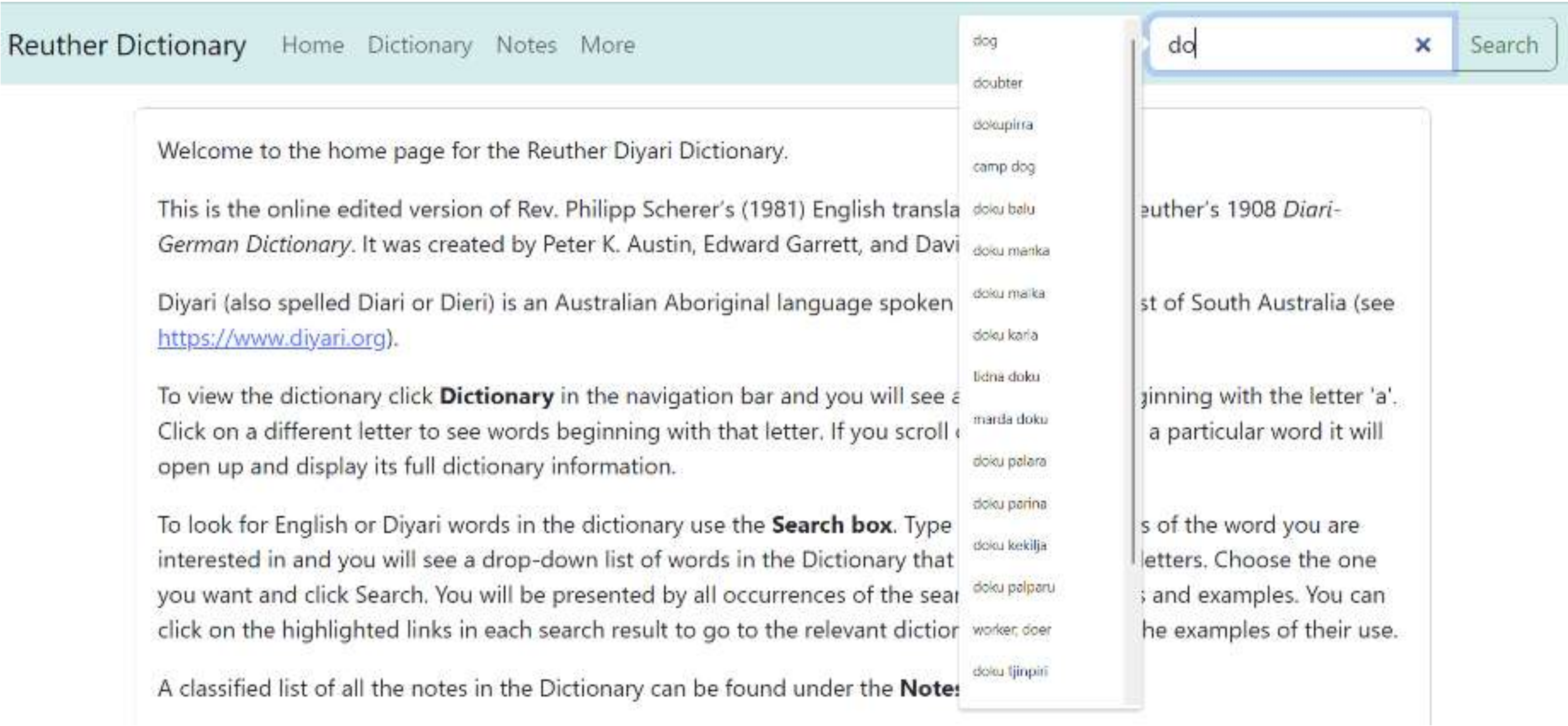
bakakaritjina n. *transformed nature*

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*

Search “do”



Welcome to the home page for the Reuther Diyari Dictionary.

This is the online edited version of Rev. Philipp Scherer's (1981) English transla euther's 1908 *Diari-*
German Dictionary. It was created by Peter K. Austin, Edward Garrett, and Davi

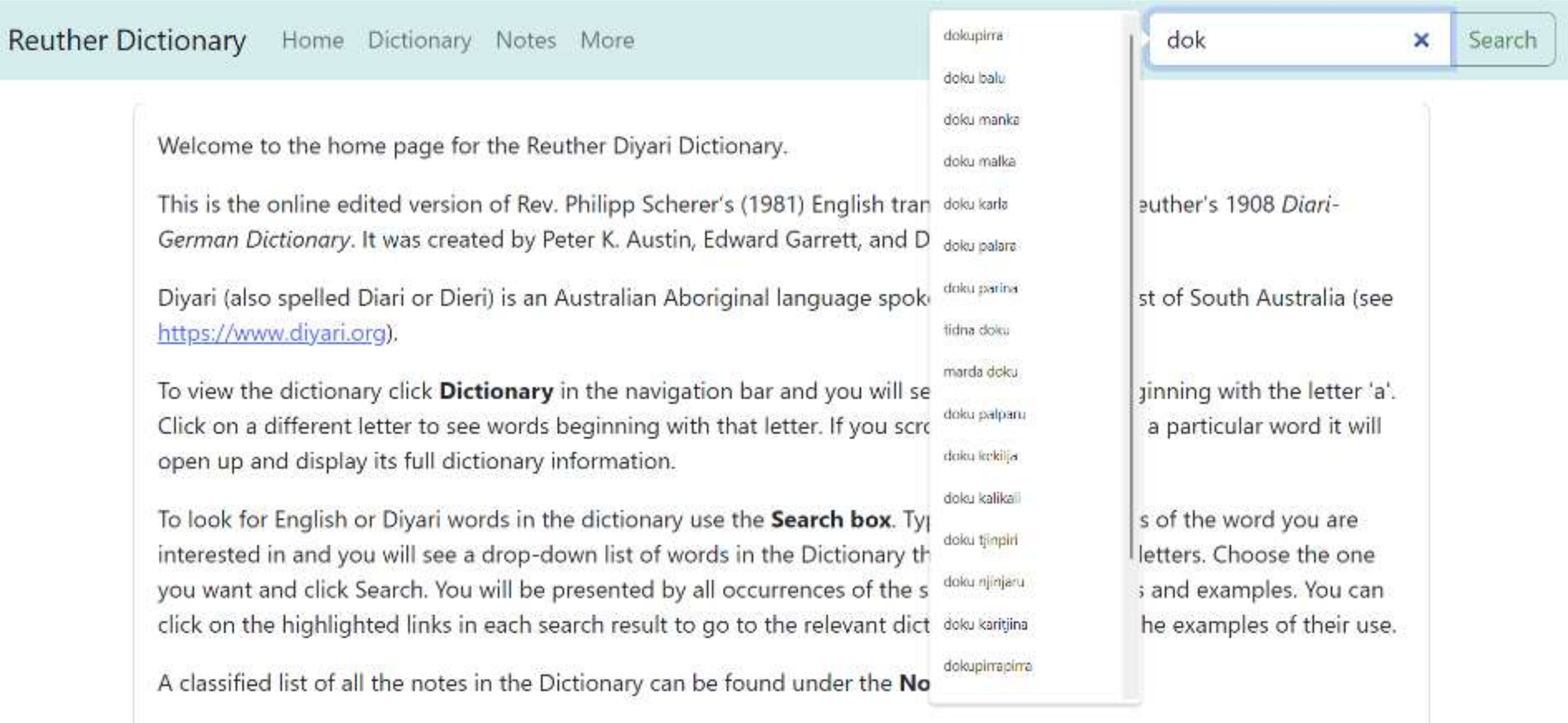
Diyari (also spelled Diari or Dieri) is an Australian Aboriginal language spoken st of South Australia (see
https://www.diyari.org).

To view the dictionary click **Dictionary** in the navigation bar and you will see a ginning with the letter 'a'.
Click on a different letter to see words beginning with that letter. If you scroll 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 s of the word you are
interested in and you will see a drop-down list of words in the Dictionary that letters. Choose the one
you want and click Search. You will be presented by all occurrences of the sear and examples. You can
click on the highlighted links in each search result to go to the relevant dictior he examples of their use.

A classified list of all the notes in the Dictionary can be found under the **Note**

Search “dok”



Welcome to the home page for the Reuther Diyari Dictionary.

This is the online edited version of Rev. Philipp Scherer's (1981) English tran euther's 1908 *Diari-German Dictionary*. It was created by Peter K. Austin, Edward Garrett, and D

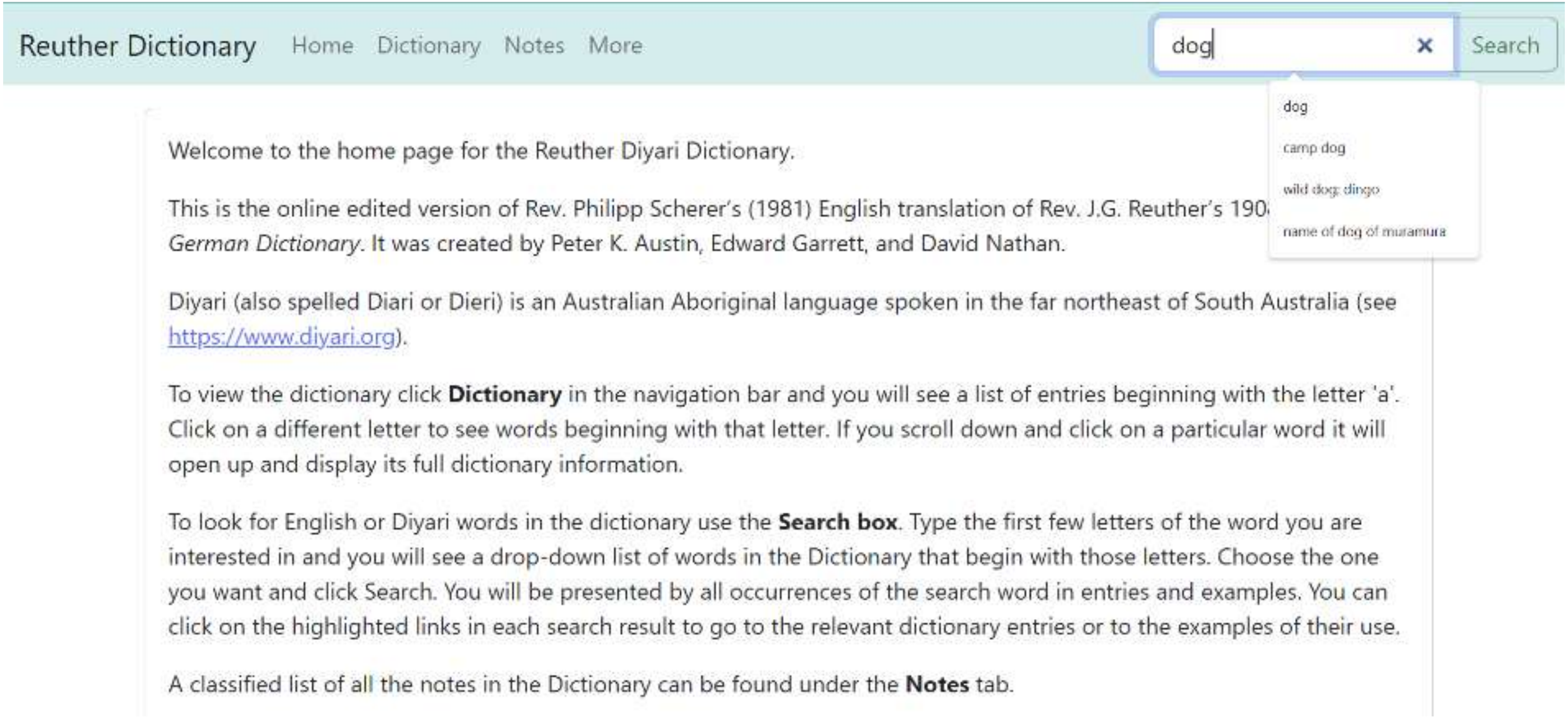
Diyari (also spelled Diari or Dieri) is an Australian Aboriginal language spok st of South Australia (see https://www.diyari.org).

To view the dictionary click **Dictionary** in the navigation bar and you will se ginning with the letter 'a'. Click on a different letter to see words beginning with that letter. If you scro 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**. Typ s of the word you are interested in and you will see a drop-down list of words in the Dictionary th letters. Choose the one you want and click Search. You will be presented by all occurrences of the s s and examples. You can click on the highlighted links in each search result to go to the relevant dict he examples of their use.

A classified list of all the notes in the Dictionary can be found under the **No**

Search "dog"



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 190 *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.

Conclusions

1. Working with legacy text materials from a language documentation perspective means dealing with often complex issues about the form, content, context and use of the original materials and analyses arising from them
2. There are many opportunities for researchers to add substantial value, especially if they can work with other historical sources and/or contemporary knowledge holders to elucidate them and the context surrounding their creation, analysis and current status
3. Maximising opportunities will require thinking about data entities, types and relationships and being explicit about them in the project design and application (e.g. in database design or XML tagging), with a very important role for metadata and meta-documentation.

Conclusions

4. Careful work with legacy text materials can also be very rewarding for researchers and communities, especially in the case of 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 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. 2017. Language documentation and legacy text materials. *Asian and African Languages and Linguistics* 11, 23-44

Austin, Peter K. 2020. Language documentation and revitalisation. In Justyna Olko & Julia Sallabank (eds.) *Revitalizing endangered languages: a practical guide*. Cambridge: Cambridge University Press.

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.

Austin, Peter K. 2025. Adding value to legacy linguistic fieldnotes: an Australian case study. Submitted to *SOAS Working Papers in Linguistics*.

Back link

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