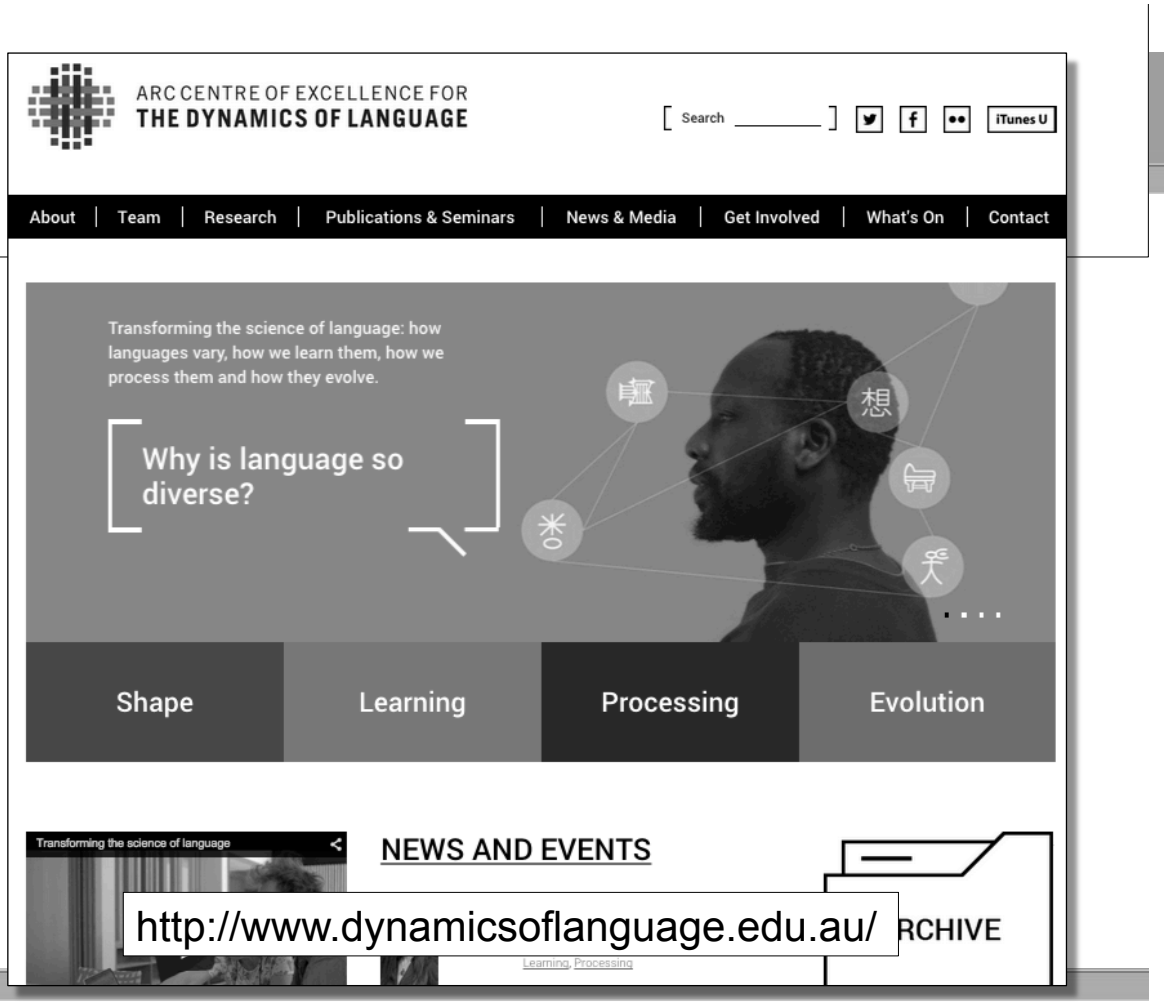
Corpora, collections, data – reusing outputs of language documentation

Nick Thieberger (University of Melbourne)

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE



CoEDL's aims

"Each program will produce varying kinds of material — recorded narratives, conversations, child language records, dictionaries, scholarly articles, statistical analyses and so on — and we aim to archive as much of it as possible. For records of small languages we will create a corpus in which it will be possible to perform the usual corpus operations (search, concordance, collocational search, frequency counts, visualisations based on well-structured data, etc.).

We will also build playable online texts (as in the current EOPAS model) that will ultimately provide stories, with interlinear text and media, for as many languages as we can arrange it for, with a map interface to make it attractive to general users."

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

Ideal corpus content types

Jan Svartvik (ed) 1990 *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund University Press. http://www.hit.uib.no/icame/london-lund/index.htm

- How does a corpus of a small language differ from other corpora?
- What is the ideal form of data for such a corpus?
- What is the most realistic form of data for such a corpus?

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

Collections and corpora of small languages

A corpus, for our purposes, is a collection of text(s) that can be searched

But primary materials that are not yet in textual format should also be brought in to the research effort, especially when there is little else available for the language:

-fieldnotes, historical manuscripts, analog recordings

Digitisation: bring analog material into current research methods

What is a corpus of a small language?

(A) Constructed in the course of fieldwork, either:
- incidental to research goal (until recently)
  - typically relies on tools to provide structure
    - if tools were used for transcription, annotation
- as the main aim of the fieldwork (especially in language documentation)
  - structured ('backslash', .eaf xml, FLEX xml)

(B) Built from sources for a particular purpose
- using tagging and conforming to a particular schema

- not 'balanced' – genres, modes, speakers, ages
- little choice about what goes in to the corpus
- everything recorded should be included

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

What is a corpus of a small language?

- Ideally
  - audio and video recorded performances
    - narratives
    - procedural texts
    - conversations
    - songs
    - etc

Actual corpora

- MS Word files on a linguist's computer, maybe intelligible to the creator (often not)
- maybe media recordings, maybe metadata
- maybe transcripts (Elan/CLAN) (XML)
- maybe a structured lexicon (Toolbox/FLEx) ('backslash, XML)
- maybe structured texts (Toolbox/FLEx)) ('backslash, XML)
- maybe social media in the language (Facebook, Twitter) (unstructured, text)

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

Why a corpus?

- What can we know about a language beyond what is given by the grammatical analysis?
- Are there citable example sentences that have context?
- Are there textual examples?
- Is there a set of records referred to by the grammar?

Corpus linguistics

- well established methods, huge literature
- corpus as a collection of texts, annotated to varying degrees
- however, little representation of ‘small’ languages in the literature

PARADISEC
Pacific and Regional Archive for Digital Sources
in Endangered Cultures
Safeguarding research
in Australia's region

Corpus linguistics

A search of the *International Journal of Corpus Linguistics* for ‘endangered’ or ‘indigenous’ or ‘Aboriginal’ or ‘First nations’ gave zero results.

Corpus linguistics

O'Keeffe, Anne, and Michael McCarthy, eds. 2010. *The Routledge Handbook of Corpus Linguistics*

'Indigenous' is not an index item, nor is 'endangered' or 'documentation', 'DoBeS', 'Corpo-AfroAs'

'Building a specialised audio-visual corpus' (93-104) by Paul Thompson

Does not mention Elan or the work of the DoBeS project, despite appearing in 2010, some 8 years after that project, with a significant emphasis on recording video and annotating multimedia, began.

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

Corpus linguistics

Ostler, Nicholas. "Corpora of Less Studied Languages."

In Lüdeling, Anke & Merja Kytö. (eds) 2008. *Corpus Linguistics: An International Handbook*. Vol. 1. Berlin ; New York: Walter de Gruyter. 457–83

Ostler: *Corpora of Less Studied Languages*

- Lists both corpora and archives
- Discusses an archive as a kind of corpus
- For the Pacific region, he lists the Aboriginal Studies Electronic Data Archive (ASEDA)† and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

- "The modern technology of corpus linguistics allows us to systematically search for particular lexical items and their collocations as well as for constructional patterns and the lexical items they accommodate, to view all findings in a concordance and to analyze the grammatical structures in their natural context." (154)

- Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Toshihide Nakayama and Keren Rice (eds) *The Art and Practice of Grammar Writing, LD&C Special Publication No.* 8. pp. 135-157.

CLUL

Welcome to CQPweb at the University of Lisbon

CORPUS QUERY PROCESSOR

Please select a corpus from the list below to enter.

Angolar FCT	Child Corpus	Reference Corpus of Contempo Portuguese (CRPC) v2.3
Fadambo FCT	Spoken Corpus Mozambique 1986-87	CRPC Portugal only v2.3
Principense FCT	Santome FCT	

Portuguese creole database
http://alfclul.clul.ul.pt/CQPweb/

Menu

Corpus queries
- Standard query
- Restricted query
- Word lookup
- Frequency lists

Corpus info
- View corpus metadata
- Corpus documentation

About CQPweb
- CQPweb main menu
- CQPweb manual
- Who did it?
- Latest news

Spoken Corpus Mozambique 1986-87: *powered by CQPweb*

Metadata for Spoken Corpus Mozambique 1986-87

Corpus name	Spoken Corpus Mozambique 1986-87
CQPweb's short handles for this corpus	moz / MOZ
Total number of corpus texts	48
Total words in all corpus texts	140,198
Word types in the corpus	7,888
Type:token ratio	0.06 types per token

Text metadata and word-level annotation

The database stores the following information for each text in the corpus:	There is no text-level metadata for this corpus.
The **primary** classification of texts is based on:	A primary classification scheme for texts has not been set.
Words in this corpus are annotated with:	lemma
	pos
The **primary** tagging scheme is:	pos
Further information about this corpus is available on the web at:	http://alfclul.clul.ul.pt/CQPweb/doc/mozambique.html

http://alfclul.clul.ul.pt/CQPweb/moz/

Let 100 flowers bloom

Recognise that corpora will be built in various ways and then try to convert results into a standard format

ANNIS - Integrating Annotations from Different Tools and Tag Sets

Chiarcos, Christian, Stefanie Dipper, Michael Balick, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. n.d. “A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets.” *Traitement Automatique Des Langues* 49 (2/2008): 217–46.

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

Bringing corpora into a standard operating environment

- Legacy material needs conversion
- We can provide templates for new research
- Train new researchers to use the templates
- Build ‘workbenches’ that sit over standard data sets to create useful search and visualisation

Using finite resources to either make many records locatable or to enrich a small set of records for automated access

- How much detail to use to annotate a set of material?
- Example of two approaches:
  – just metadata to make page images locatable
  – textual content to make it all searchable

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

Presenting fieldnotes as images

- Case study: Arthur Capell's papers >14,000 pages
  - Paper originals stored in the National Library of Australia
  - Simple means of providing access to unique resources
  - No time or funding to encode the text of the images
- Metadata created by the executor of the estate
- Photographic images (not scanned)
  - Portable photographic rig, images taken in the executor's house
  - Rapid and easy capture

Pacific and Regional Archive for Digital Sources in Endangered Cultures

PARADISEC

Home
Provenance
Series List
Search

Arthur Capell (1902-1986), Guide to Records

Listed by Peter Newton, Austehc and PARADISEC

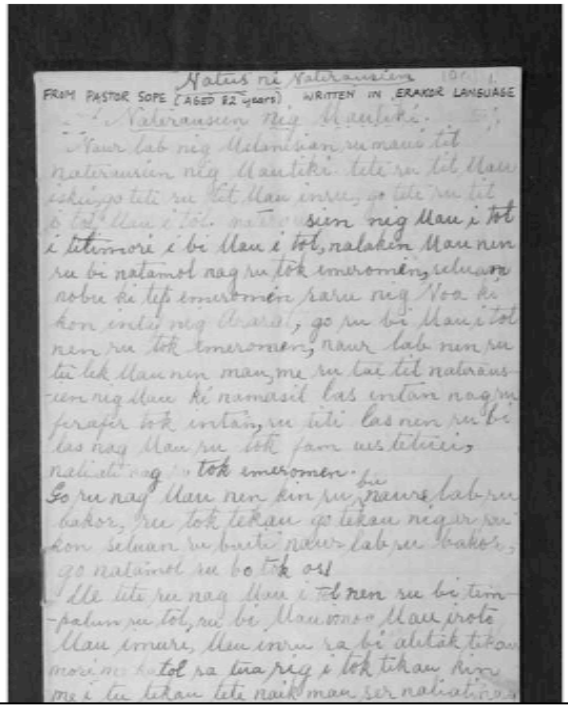
Arthur Capell was an Australian linguist and ethnographer who spent much time recording and documenting both Australian Aboriginal languages and endangered languages in the Asia - Pacific region. The Arthur Capell textual collection consists of some thirty boxes of notes, transcripts, and other supporting materials. What is documented in this guide are some of the records relating to his non-Australian linguistic work that have been digitally imaged.

Please do not copy material from this site for further distribution but rather link to this site. PARADISEC has raised funds to digitise this collection and would like to be recognised for the work that we have put into developing the online presentation of fieldnotes. If you copy and distribute this data and do not acknowledge PARADISEC's work then we will have to put password protection on the data.

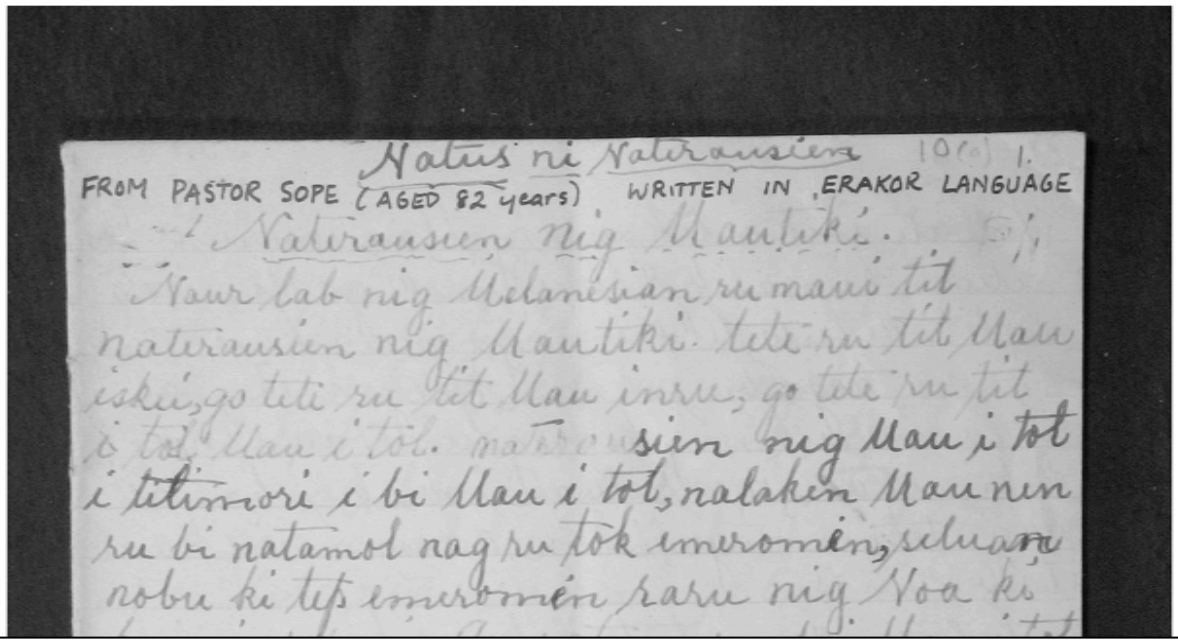
[Details]

Arthur Capell (1902-1986), Guide to Records
Item: VEFAT25
Contact PARADISEC
To print this page you may need to select landscape paper orientation for better results.

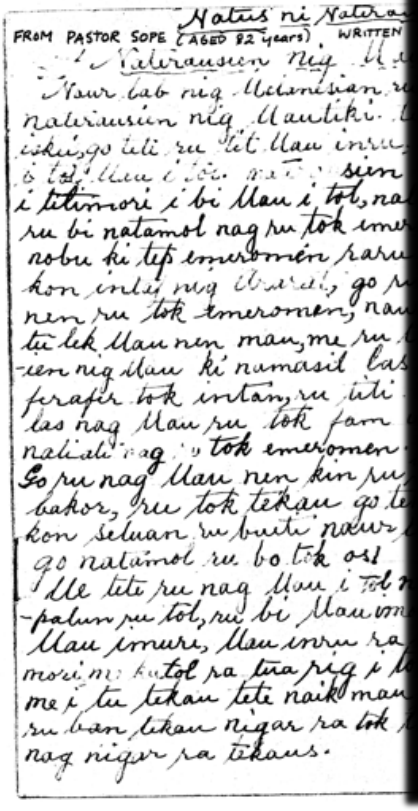
Read / Enlarge <-- Image 1 of 30 --> Back to Guide



http://paradisec.org.au/fieldnotes/AC2.htm



http://paradisec.org.au/fieldnotes/AC2.htm



(1) Natirausien nig Mautiki

1. Naur lab nig Melanesian ru maui til naterausien nig Mautiki tete ru til Mau iskei, go tete ru til Mau inru, go tete ru til i tok Mau i tol naterausien nig Mau i tol i tilimori i bi Mau i tol, nalaken Mau nen ru bi natamol nag ru tok emeromen, seluan nobu ki tip emeromen raru nig Noa ki kon intaf nig Ararat, go ru bi Mau i tol nen ru tok emeromen, naur lab nen ru tu lek Mau nen mau, me ru tae til naterausien nig Mau ki namasil las intan nagru ferafer tok intan, ru tili las nen ri bi las nag Mau ru tok fam uis tetuei, naliati nag ru tok emeromen.

2. Go ru nag Mau nen kin ru bu naure lab ru bakor, ru tok tekau go tekau nigar ru kon seluan ru bueti naur lab ru bakor, go natamol ru bo tok os!

3. Me tete ru nag Mau i tol nen ru bi tempalun ru tol, ru bi Mau imoo Mau iroto Mau imuri, Mau inru ra bi atlak tikau mori me katol ra tua prig i tok tikau kin me i tu tkau tete naik mau ser naliatinag ru ban tekau nigar

(1) Natrauswen ni Mautiki

1. Naur lap ni Melanesia rumaui til natrauswen ni Mau. Tete ru til Mau iskei, tete rutil inru, go tete rutil Mau itol. Natrauswen ni Mau itol itilmori kin ipi Mau itol, nlaken Mau nen rupu natañol nag ru tok emeromen. Selwan nab itap emermen go raru ni Noa ikon ntaf ni Ararat, naur lap rutap lek Mau nen mau me rutae til natrauswen ni Mau ki nmasil las nag rufrafer tok ntan, rutli nag las nen rupi las nag Mau ruto fam wes tetuei, malnen ruto emermen.

2. Go runag Mau nen kin rupo naur lap rupakor. Ruto pan tkau go tkau negar rukon go selwan rupueti naur lap rupakor, go natañol rupo tkos.

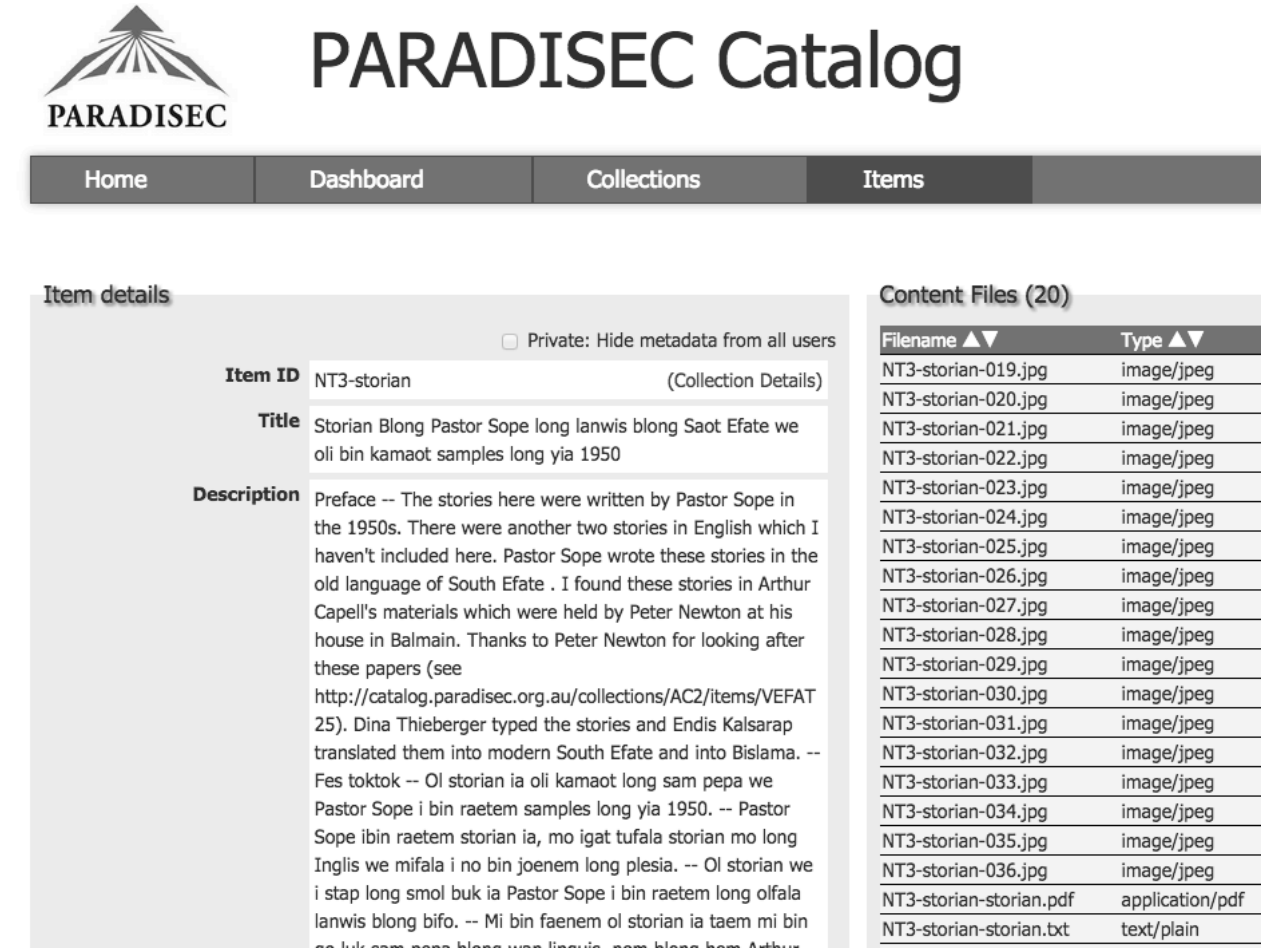
3. Me tete runag Mau tol nen rupi temp alun, rupi Mau Imoo, Mau Iroto, ga Mau Imuri. Mau inru rapi atlak tkau mori me katol ratua tkau gar ito tkau kin me naliati itap tkau tete naik mau ser naliati nag rupan tkau gar rato tua, tete naik nag gar ratkaus.

(1) Storian blong Mauitiki

1. Fulap aelan blong Melanesian oli talem stori blong strong man blong faet. Sam oli talem se igat wan, sam oli talem igat tu ma sam oli talem se igat tri. Stori blong trifala strong man blong faet hemi tru from ol man blong faet ya oli ol man we oli stap long wol. Taem we wota i bin drae long wol mo Bot blong Noa i fas long hil blong Ararat, fulap aelans oli no luk ol man blong faet ya be oli save talem stori blong hem from ol pispis klas we oli stap olbaot oli se hemi ol kap we ol man ya oli stap kakae long hem bifo taem oli stap long wol.

2. Mo oli talem se ol strong man blong faet ya nao oli pulum ol plante aelan oli kamaot. Oli stap go huk mo huk blong olgeta i fas mo taem oli pulum, plante aelan oli kamaot, mo ol man oli stap long hem.

3. Be sam oli talem se ol trifala strong man blong faet ya oli brata, ol nem blong olgeta oli Imoo, Iroto mo Imuri. Tu long trifala strong man ya tufala igat huk be tufala i givim huk blong tufala



One-off examples of encoded text for a small language

Web-accessible Apache language linguistic database and text archive, a republication of Hoijer's 1938 monograph originally published by the University of Chicago.

http://etext.lib.virginia.edu/apache/ChiMesc2.html (2008)

now resolves to ->

http://xtf.lib.virginia.edu/xtf/view?docId=Apache/uvaGenText/tei/Chi01.xml

1. 55 Apache language texts (in Chiricahua and Mescalero dialects) representing a variety of genres: stories, songs, prayers and speeches, performed by nine different Apache speakers

2. English translations of the Apache texts, developed by Hoijer in consultation with Apache speakers

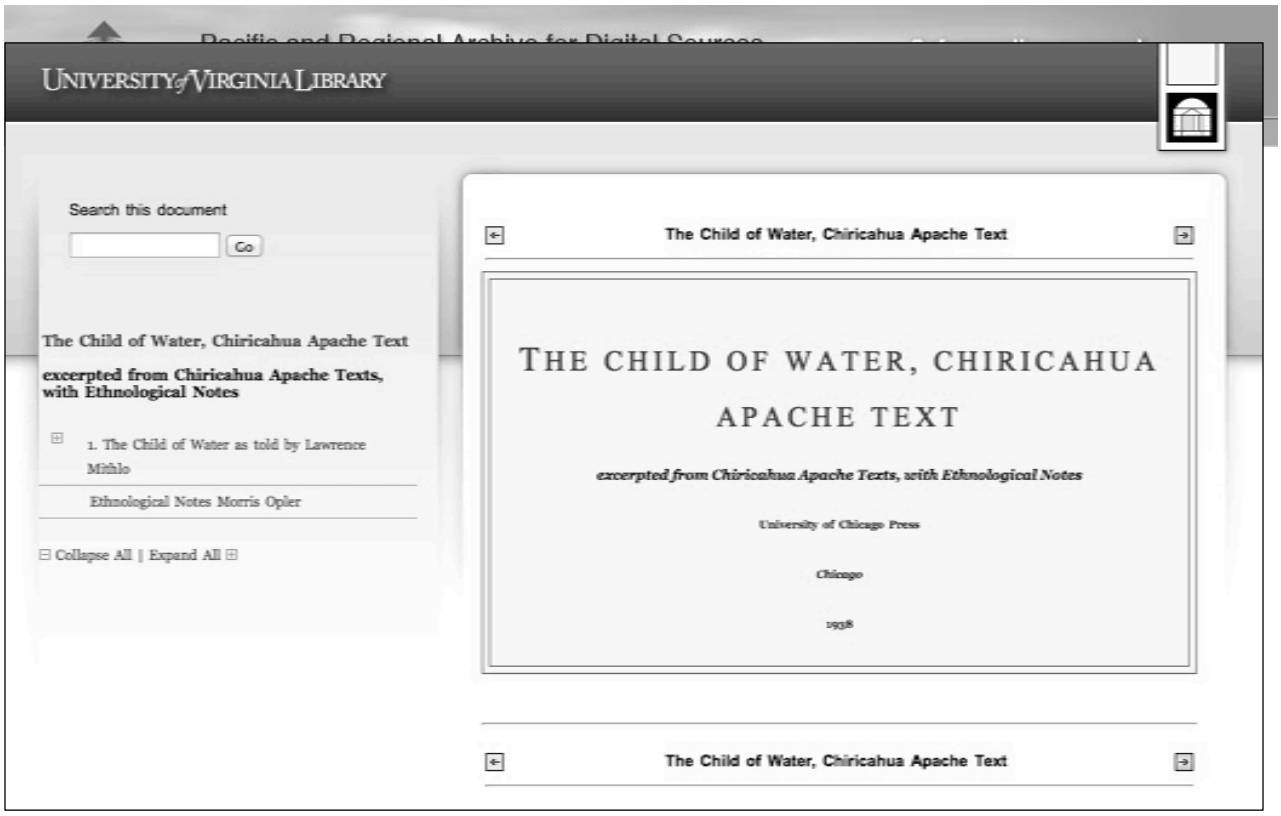
3. Ethnological notes to the texts, contributed by Morris Opler, a cultural anthropologist

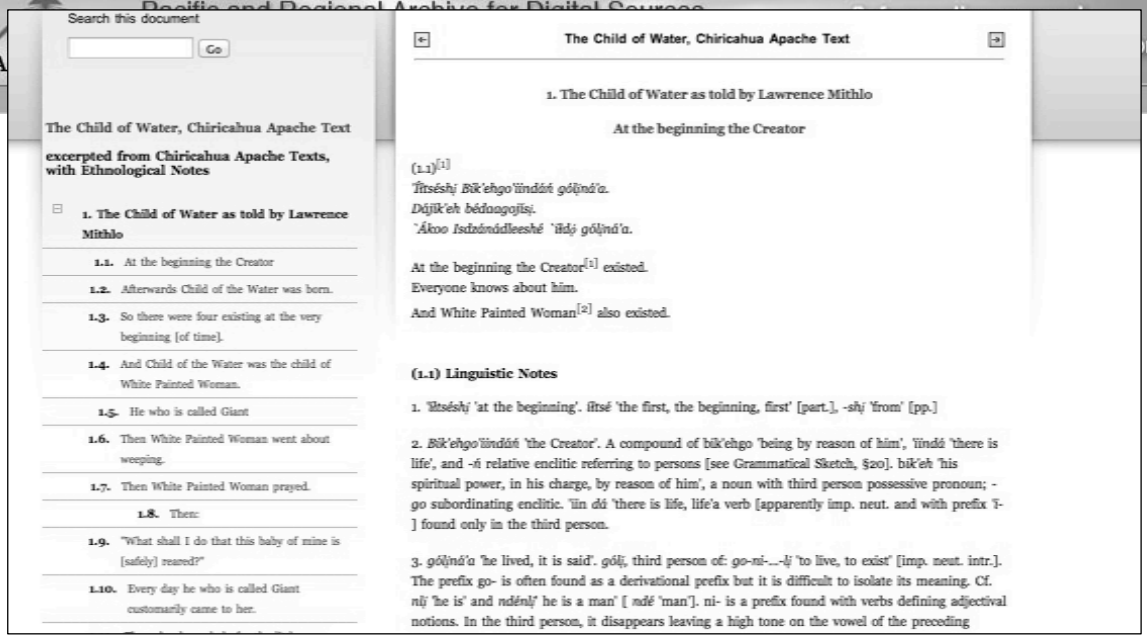
4. Hoijer's linguistic notes, containing morphological analysis of words and phrases as they appear in the texts

5. Hoijer's grammatical outline of Chiricahua and Mescalero Apache language varieties

6. a search-engine for Apache and English language searches

7. a lexicon to the texts developed from Hoijer's linguistic analysis





Wide shallow or narrow deep

- Where effort is being made to annotate and enrich a corpus that should then be usable within a cross-corpus search
- Where the effort is to make primary material available, that too should be available, even at the level of metadata
  - with an annotation framework that permits it to be enriched

EOPAS – text and media corpus presentation

Upload text in interlinear format

Upload media

Present text and media online

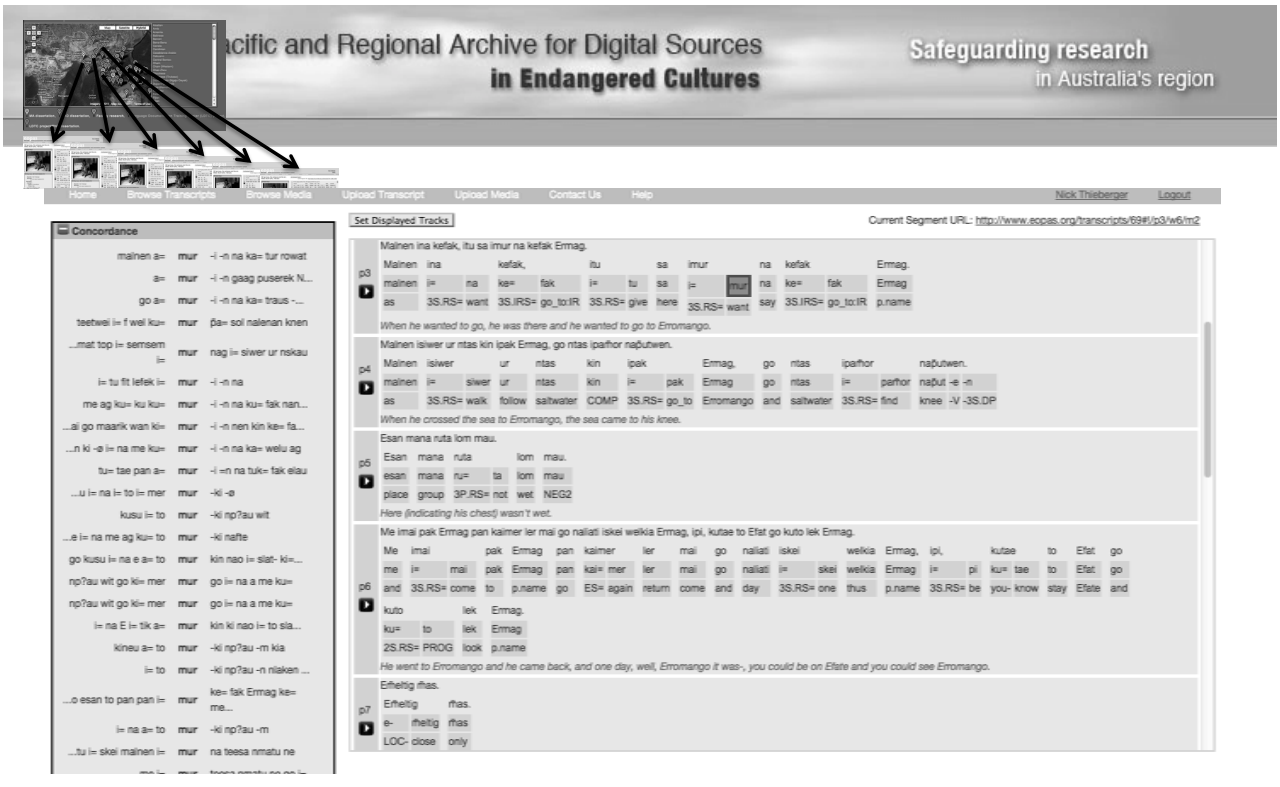
Citable text and media (to the level of the morpheme)

- Planned

  Allow media in PARADISEC to be called directly

  Provide online annotation

http://www.eopas.org/transcripts/55



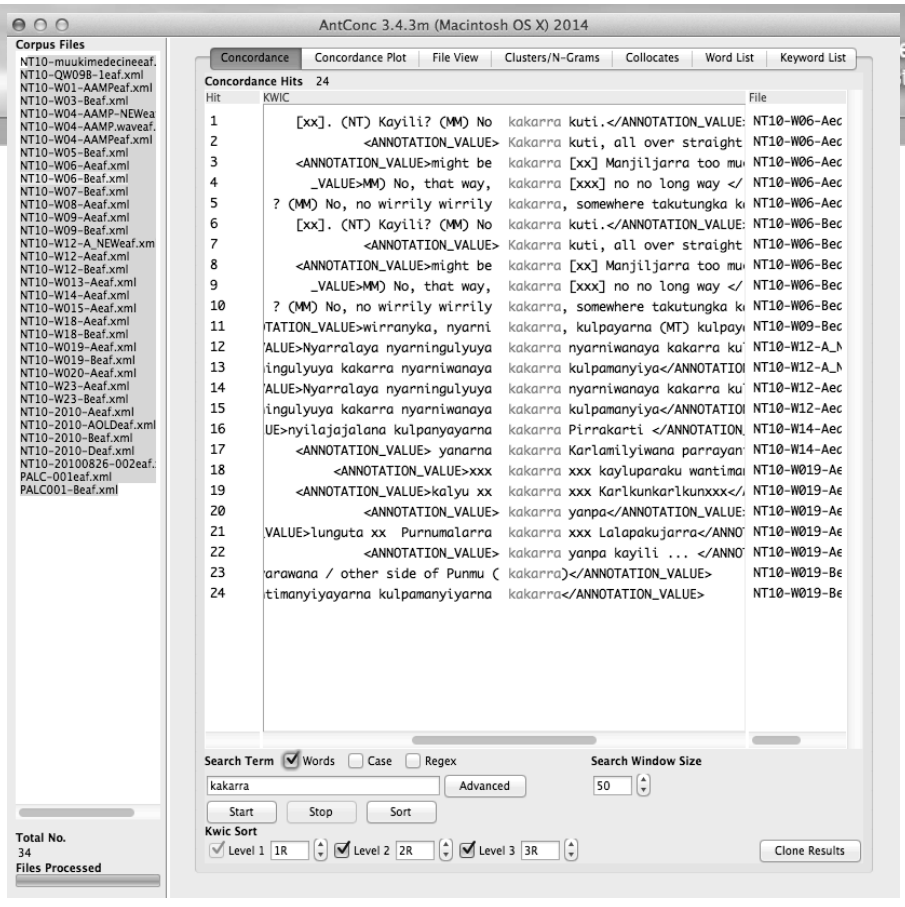
http://www.eopas.org/transcripts/69#!/p3/w6/m2

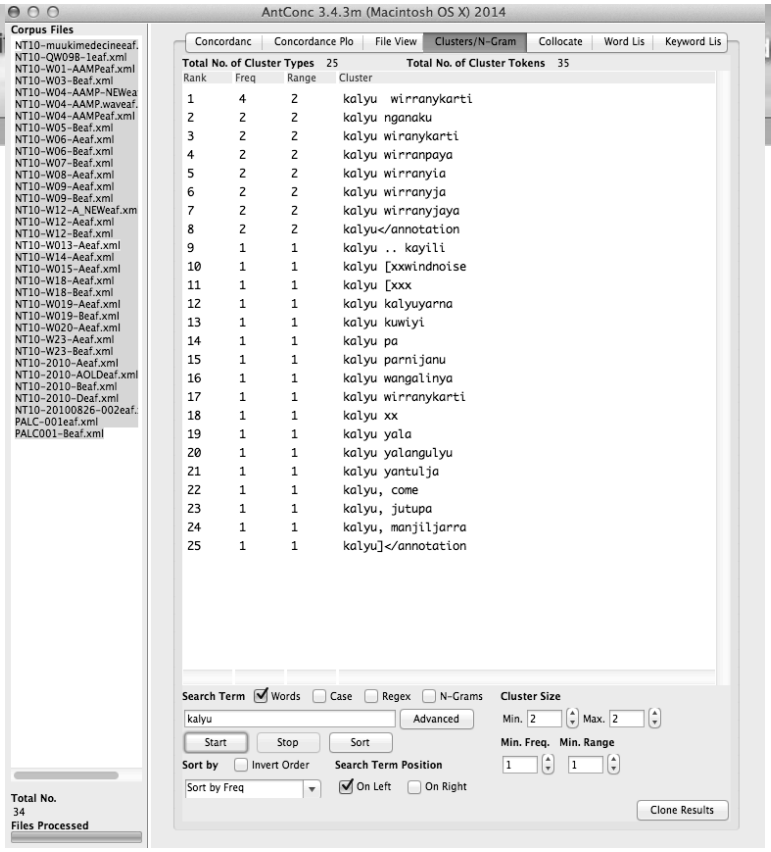
Using AntConc to query small corpora

- AntConc – won't open .eaf files
- Knows about xml, so rename .eaf as .xml

http://www.laurenceanthony.net/software/antconc/

PARADISEC




AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
NT10-muukimedecineeaf.
NT10-QW09B-1eaf.xml
NT10-W01-AAMPeaf.xml
NT10-W03-Beaf.xml
NT10-W04-AAMP-NEWea
NT10-W04-AAMP.waveaf.
NT10-W04-AAMPeaf.xml
NT10-W05-Beaf.xml
NT10-W06-Aeaf.xml
NT10-W06-Beaf.xml
NT10-W07-Beaf.xml
NT10-W08-Aeaf.xml
NT10-W09-Aeaf.xml
NT10-W09-Beaf.xml
NT10-W12-A_NEWeaf.xm
NT10-W12-Aeaf.xml
NT10-W12-Beaf.xml
NT10-W013-Aeaf.xml
NT10-W14-Aeaf.xml
NT10-W015-Aeaf.xml
NT10-W18-Aeaf.xml
NT10-W18-Beaf.xml
NT10-W019-Aeaf.xml
NT10-W019-Beaf.xml
NT10-W020-Aeaf.xml
NT10-W23-Aeaf.xml
NT10-W23-Beaf.xml
NT10-2010-Aeaf.xml
NT10-2010-AOLDeaf.xml
NT10-2010-Beaf.xml
NT10-2010-Deaf.xml
NT10-20100826-002eaf.
PALC-001eaf.xml
PALC001-Beaf.xml
Total No.
34
Files Processed
Concordanc
Concordance Plo
File View
Clusters/N-Gram
Collocate
Word Lis
Keyword Lis
Total No. of Cluster Types 25
Total No. of Cluster Tokens 35
Rank Freq Range Cluster
1 4 2 kalyu wirranykarti
2 2 2 kalyu nganaku
3 2 2 kalyu wiranykarti
4 2 2 kalyu wirranpaya
5 2 2 kalyu wirranyia
6 2 2 kalyu wirranyja
7 2 2 kalyu wirranyjaya
8 2 2 kalyu</annotation
9 1 1 kalyu .. kayili
10 1 1 kalyu [xxwindnoise
11 1 1 kalyu [xxx
12 1 1 kalyu kalyuyarna
13 1 1 kalyu kuwiyi
14 1 1 kalyu pa
15 1 1 kalyu parnijanu
16 1 1 kalyu wangalinya
17 1 1 kalyu wirranykarti
18 1 1 kalyu xx
19 1 1 kalyu yala
20 1 1 kalyu yalangulyu
21 1 1 kalyu yantulja
22 1 1 kalyu, come
23 1 1 kalyu, jutupa
24 1 1 kalyu, manjiljarra
25 1 1 kalyu]</annotation
Search Term Words Case Regex N-Grams
kalyu
Advanced
Start
Stop
Sort
Sort by Invert Order
Sort by Freq
Search Term Position
On Left
On Right
Cluster Size
Min. 2 Max. 2
Min. Freq. Min. Range
1 1
Clone Results


AntConc 3.4.3m (Macintosh OS X) 2014
Corpus Files
NT10-muukimedecineeaf.
NT10-QW09B-1eaf.xml
NT10-W01-AAMPeaf.xml
NT10-W03-Beaf.xml
NT10-W04-AAMP-NEWea
NT10-W04-AAMP.waveaf.
NT10-W04-AAMPeaf.xml
NT10-W05-Beaf.xml
NT10-W06-Aeaf.xml
NT10-W06-Beaf.xml
NT10-W07-Beaf.xml
NT10-W08-Aeaf.xml
NT10-W09-Aeaf.xml
NT10-W09-Beaf.xml
NT10-W12-A_NEWeaf.xm
NT10-W12-Aeaf.xml
NT10-W12-Beaf.xml
NT10-W013-Aeaf.xml
NT10-W14-Aeaf.xml
NT10-W015-Aeaf.xml
NT10-W18-Aeaf.xml
NT10-W18-Beaf.xml
NT10-W019-Aeaf.xml
NT10-W019-Beaf.xml
NT10-W020-Aeaf.xml
NT10-W23-Aeaf.xml
NT10-W23-Beaf.xml
NT10-2010-Aeaf.xml
NT10-2010-AOLDeaf.xml
NT10-2010-Beaf.xml
NT10-2010-Deaf.xml
NT10-20100826-002eaf.
PALC-001eaf.xml
PALC001-Beaf.xml
Heath-Text13.1.1-13.3.3.
Wubuy-Lexicon.xml
Total No.
36
Files Processed
Concordance
Concordance Plo
File View
Clusters/N-Gram
Collocates
Word Lis
Keyword Lis
File View Hits 171
File Wubuy-Lexicon.xml
</AStr>
</LexEntry_HeadWord>
<LexEntry_Senses>
<LexSense number="1" id="hvo3753">
<MoMorphSynAnalysisLink_MLPartOfSpeech>
<AStr ws="en">
<Run ws="en">n</Run>
</AStr>
</MoMorphSynAnalysisLink_MLPartOfSpeech>
<MoMorphSynAnalysisLink_MLInflectionClass>
<AStr ws="en">
<Run ws="en">fem</Run>
</AStr>
</MoMorphSynAnalysisLink_MLInflectionClass>
<LexSense_Definition>
<AStr ws="en">
<Run ws="en">hawksbill turtle, Eretmochelys imbricata
(has large, parrotlike beak Syn: garruba. , a fly sp. (uncommon
sense).</Run>
</AStr>
</LexSense_Definition>
</LexSense>
</LexEntry_Senses>
Search Term Words Case Regex
fem
Advanced
Start
Stop
Hit Location
9
Clone Results

Conclusions

- Corpora of small languages typically vary considerably
- Best to address each in its own terms
- Build methods to operate with this diversity
- Provide templates for new researchers

PARADISEC
Pacific and Regional Archive for Digital Sources in Endangered Cultures
Safeguarding research in Australia's region

References

Chiarcos, Christian, Stefanie Dipper, Michael Balick, Ulf Leser, Anke Lüdeling, Julia Ritz, and Manfred Stede. n.d. "A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets." *Traitement Automatique Des Langues* 49 (2/2008): 217–46.

Himmelmann, Nikolaus P. 2012. Linguistic Data Types and the Interface between Language Documentation and Description. *Language Documentation & Conservation* 6. 187-207.

Lüdeling, Anke, and Merja Kytö, eds. 2008. *Corpus Linguistics: An International Handbook*. Vol. 1. 2 vols. W. de Gruyter.

Lüdeling, Anke, and Merja Kytö, eds. 2009. *Corpus Linguistics: An International Handbook*. Vol. 2. 2 vols. Handbücher Zur Sprach- Und Kommunikationswissenschaft = Handbooks of Linguistics and Communication Science, Bd. 29.1-. Berlin; New York: Walter de Gruyter.

Mosel, Ulrike. 2014. Corpus linguistic and documentary approaches in writing a grammar of a previously undescribed language. In Toshihide Nakayama and Keren Rice (eds) The Art and Practice of Grammar Writing, LD&C Special Publication No. 8. pp. 135-157.

Paul Thompson. 2010. Building a specialised audio-visual corpus. In . O'Keeffe, Anne, and Michael McCarthy, eds. *The Routledge Handbook of Corpus Linguistics*. 1st ed. 93-104. Routledge Handbooks in Applied Linguistics. New York, N.Y: Routledge.

Woodbury, Anthony. 2014. Archives and audiences: toward making endangered language documentations people can read, use, understand, and admire. In David Nathan and Peter Austin (eds.) *Language Documentation and Description, Volume 12: Special Issue on Language Documentation and Archiving*. London: SOAS. 19-36.

Wynne, Martin, ed. 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oakville, CT: Oxbow Books.

Australian Research Council – LIEF grants LE110100142, LE0560711, LE0453247
ARC DP0450342, DP0984419, FT140100214

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE