Corpora, collections, data – reusing outputs of language documentation

Nick Thieberger (University of Melbourne)
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.”

Ideal corpus content types

• 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?

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 into 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

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/FLEex) (‘backslash, XML)
- maybe structured texts (Toolbox/FLEex) ) (‘backslash, XML)
- maybe social media in the language (Facebook, Twitter) (unstructured, text)

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

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.

Corpus linguistics

Ostler, Nicholas. “Corpora of Less Studied Languages.”

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)

“‘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)

Welcome to CQPweb at the University of Lisbon
CORPUS QUERY PROCESSOR

Please select a corpus from the list below to enter.

<table>
<thead>
<tr>
<th>Angolafct...</th>
<th>Child Corpus</th>
<th>Reference Corpus of Contemporary Portuguese (CRPC) v2.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fadambofct...</td>
<td>Spoken Corpus Mozambique 1986-87</td>
<td>CRPC Portugal only v2.3</td>
</tr>
<tr>
<td>Principensefct...</td>
<td>Santomefct...</td>
<td></td>
</tr>
</tbody>
</table>

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

---

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

### Metadata for Spoken Corpus Mozambique 1986-87

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

### 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
- pos

The **primary** tagging scheme is:

Further information about this corpus is available on the web at:


---

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


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

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
Arthur Capell (1902-1986), Guide to Records

Listed by Peter Newton, Austhec 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.

Copyright: PARADISEC believes that many of the items provided through this guide are no longer the subject of copyright restrictions, or have been donated for display in this service by the Copyright owners. However, PARADISEC invites any individuals who believe they hold current rights over items provided through this service to make contact.

[Details]

- About the records
  - Scope and content
  - How to use this finding aid
  - Archival terms

- Provenance - creators and former custodians

- Series list and summary descriptions

- Inventory listing by series
  - Series 01 - Personal and Biographical Material
  - Series 02 - General Linguistic and Ethnological Materials
  - Series 03 - Indonesia - Regions and Languages
  - Series 04 - Bougainville, Buka and Offshore Islands - Regions and Languages
  - Series 05 - Melanesia and Oceania - Regions and Languages
  - Series 06 - Solomon Islands - Regions and Languages
  - Series 07 - Timor - Regions and Languages
  - Series 08 - Vanuatu - Regions and Languages
  - Series 09 - Fiji - Regions and Languages

Naturausen ni Mautuki


1. Pulup amane blong Melanesian oli talem stori blong strong man blong faet. Sam olim talem se iag wan, sam olim talem iag tu ma sam olim talem se iag tu. Stori blong tufala strong man blong faet herna tru from olim man blong faet yu el olim man we olim stap long we. Dae we wota in din drag long we mo Bot blong Noa i fas long hil blong Ararat. Pulup amane olim no huk el olim blong faet yu el olim save talem stori blong herna from olim pipsis kla we olim stap olbom olim se herna el kap we el olim man el olim stap kain long we. Hem hif olim taen el olim stap long we.

2. Me talem se el strong man blong faet yu el olim pulum olim pinta amane olim karnait. Olim stap go huk mo huk blong olbega i fas mo taen olim pulum, pinta amane olim karnait, mo olim el olim stap long hem.

3. Be sam olim talem se el tufala strong man blong faet yu el olbuma, olim nobl olim olbega olim imuro. Olim stap go 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.


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
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/
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

References


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

ARC CENTRE OF EXCELLENCE FOR
THE DYNAMICS OF LANGUAGE