FieldLing seminar, 3rd September 2024

Corpus creation, archiving and use

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

www.peterkaustin.com

General issues for discussion

1. Collecting and analysing language research materials (creating a corpus)
2. What do we mean by ‘language documentation’, ‘language description’, and ‘language revitalisation’?
3. How do you analyse a corpus?
4. How do you archive a corpus?
5. How can a corpus be used? – for description, revitalisation, other purposes? (see also Seifart presentation)

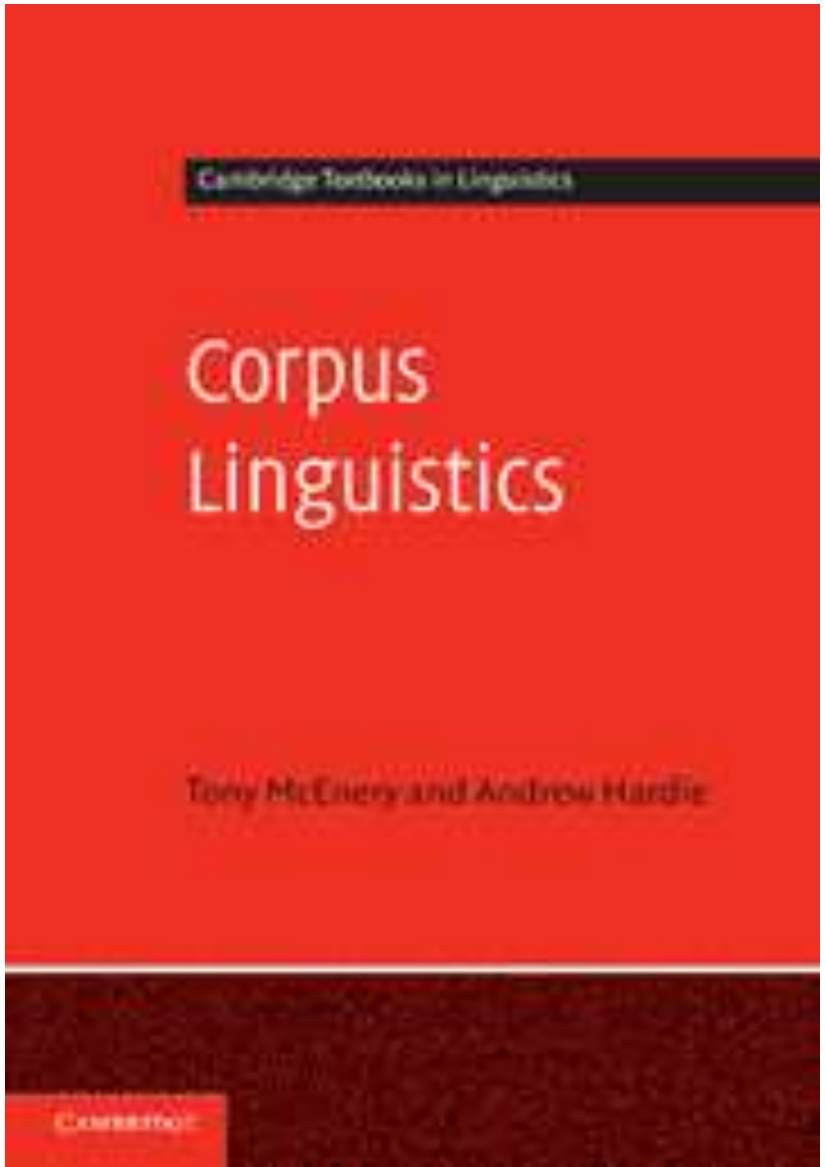
What is a corpus (plural corpora)?

Traditional definition: A collection of linguistic data, either compiled as written texts or as a transcription of recorded speech. The main purpose of a corpus is to verify a hypothesis about language - for example, to determine how the usage of a particular sound, word, or syntactic construction varies. Corpus linguistics deals with the principles and practice of using corpora in language study. A computer corpus is a large body of machine-readable texts.

(Crystal, David. 1992. *An Encyclopedic Dictionary of Language and Languages*. Oxford: Blackwell.)

Note:
1. emphasis on written text;
2. primarily quantitative analysis;
3. software tools;
4. particular collection is justified by research hypothesis (goals)

A good introduction



Tony McEnnery & Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory, and Practice*. Cambridge: Cambridge University Press.

How do you create a corpus?

Traditional approach: collect a set of texts (already born digital or by scanning print publications):

1. a sample or collection which is **representative** with regards to the research hypothesis;
2. there may be an attempt to **balance** the corpus to represent various non-linguistic variables (e.g. x% novels, y% poetry, z% conversation)
3. with a defined size (e.g. 1 million words), and content (e.g. English literature).
4. e.g. British National Corpus (1980-90, 1 million words, multiple written genres) https://www.english-corpora.org/bnc/
5. e.g. frTenTen: Corpus of French Web (current, 10 billion words, European Canadian, and African French) https://www.sketchengine.eu/frtenten-french-corpus/

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

For discussion

1. Why do we want to describe languages?

2. Give some examples of language description

3. Who is the audience for language description?

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)

A new approach: 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” (Himmelmann 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
- Woodbury (2003, 2011) ‘transparent records of a language’

What's new in documentation?

- ***Data focus*** – this is Himmelman's "primary data" and includes **audio**, **video**, **still images**, and **text**, but also **structured data** derived from processed materials (transcribed, translated, annotated digital files). A collection of such material is called a **corpus**. See Himmelmann 2012. We discuss problems with this later.
- ***Accountability*** – we expect the materials ("primary" and analysed) to be made available to others. Some have argued for **reproducibility**, i.e. the possibility of recreating the researcher's analytical steps to see if the outcome is the same (or different). See Berez-Kroeker et al 2017. We discuss problems with this later.
- ***Preservation*** -- long-term storage in safe archival facilities where the data and analysis (corpora) can be safeguarded for the long term (including refreshing data formats to take into account changing software)
- ***New software tools*** – data and analysis is stored in digital files and access is mediated via computer software (e.g. Praat, ELAN, FLEx)

For discussion

- Why do we want to document languages/dialects by collecting instances of language performance (use in social-cultural-political-economic context)?

- If languages/dialects are disappearing, what is the point of studying them?

Metadata and meta-documentation

In order to organise, manage, understand, and analyse the corpus we need:

- **Metadata –** data about the data – several types at file-level (see Nathan & Austin 2004):
  - *cataloguing* — title, speakers, collectors, time and place of recording, language name etc.
  - *descriptive* — information about content, relationship to other resources etc.
  - *structural* — what structural devices and patterns exist in the document etc.
  - *technical* — performance and preservation information, description of formats etc.
  - *administrative* — work log, responsibilities, access protocol statements etc.
- **Meta-documentation –** metadata at the project level: goals, corpus theory, data collection and analytical methods, stakeholders, ethics (informed consent), access and use

This is also important for archiving (see below)

Data collection methods

1. **Elicitation** (interviewing):
   - Translation (L1→ L2, L2 → L1)
   - Speaker judgements: “can you say xxxxxx?”, “does yyy sound/mean the same as xxx?”

2. **Narratives** (telling stories)

3. **Conversation** (2 or more participants)

4. **Experimentation** – puzzles, games, other tasks (video description, image description = use of “stimuli”)

5. **Participant observation** – spending time with speakers observing language use and attempting to use the language oneself

Each method has strengths and weaknesses

Metadata collection and management

1. Can be done manually (pen and paper) but best done electronically for ease of storage, searching, sharing, and restructuring
2. Plain text files, spreadsheet (Excel), or database (Access, MySQL)
3. Dedicated metadata software – SayMore (stand alone), CIMDI maker (online) etc.

Archives may have preferences on metadata tools and formats – check when you are first designing your project and planning your corpus

Corpus management

Two basic principles to manage your corpus:

1. Develop a **file naming system** and stick to it rigorously:
   - Use only ASCII symbols (English letters and numbers, **no punctuation**, **no spaces**, no accented characters) – if necessary use – and _
   - Names should contain one period (.) only followed by an extension (3-4 characters) giving the file type: .docx, .txt, .wav, .jpg, .eaf
   - Keep names short and do not try to stuff metadata into file names
   - Make names sortable, e.g. using ISO date standard
   - e.g. 2021-09-09_FieldLing_corpus.pptx
2. Develop a **folder naming system** that makes sense to you and put files in their correct place so you can always find them

Folders

Example of my system:

Talks
- 2019
- 2020
- 2021
  - 2021-05-28_Pakistan
  - 2021-06-24_NEIndia
  - 2021-09-09_Paris
    - 2021-09-09_FieldLing_corpus.pptx

Research
- Diyari
- Jiwarli

And most importantly

Three further principles:

BACKUP

BACKUP

BACKUP

1. Make multiple copies of your corpus in multiple formats (HD, SSD, Cloud) in multiple locations on a regular basis (LOCKSS)
2. Consider continuous backup (e.g. Carbonite) or regular automatic scheduled backup
3. You **will** lose data and analysis – your job is to make the loss the smallest possible

Secondary corpora (‘legacy materials’)

- 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

Secondary corpora example

- Peter Austin *Toolbox* databases (lexicon, glossed and annotated texts) from S.A. Wurm's 1955-57 handwritten notes of Australian languages
- Retype original, add metadata on sources (speaker, recorder, fieldnotes location), abbreviation definitions, date of last edit
- Add sentence analysis: phonemicization, morpheme glossing, part of speech, free translation in English, notes, link to lexicon (lexnum), link to abbreviations
- Add lexicon: headword, gloss, definition, scientific name, scientific name source, picture, semantic relations (synonym, antonym, cf), notes, cognates, example sentence link (text, free translation)

(Elpi Parlow)	(MacFarlane) (R) (M.Q.)
wannju	i:nipi punju- ε: baddja layambuni
	e.g. ~ pl. ze, 'e co / bad-ond,- 'e co)
aε:, wa	la dó daddja layambunji i:nipi punjugu
	ze'o / co
minjpu	i:nipi punju ze:z
banimaina	leonhim
wanimma	gagulinu daddja lambunna
buppani	punju or, 'ze ?a 6/fn /co.
punju ze	
i:nipida	jalida (punju 2 'ze lo 2o

Toolbox - malyangapa_lexicon

File Edit Database Project Tools Checks View Window Help

[no filter]

malyangapa_lexicon

\lx	**dhadja**
\a	
\u	
\lxnum	029
\lg	MI
\ps	*vtr*
\ge	bite
\def	bite
\sd	Actions
\eth	
\sci	
\sci_source	
\nt	
\syn	
\ant	
\cf	**dhaba**
\se	
\lgcf	
\ety	
\rec	SW
\sp	HQ
\x_ref	011
\xv	**dhadjangarndambunyi gunyungu**
\xe	The dog bit me
\wnum	425 426 427 428 429 430 431

malyangapa_notes

\snum	017
\text	**wanyu yinigi gunyuyi, dhadjalangambuni**
\morpheme	*wanyu yinigi gunyu -yi dhadja -langa -mbu -ni*
\gloss	bad that dog -emph bite -might -3sg.A -2sg.P
\ps	n dem n -suff vtr -suff -suff -suff
\lxnum	028 035 002 -034 029 -049 -031 -050
\free_translation	This dog is bad, it will bite you
\reference	SW1/2As01
\recorder	SW
\speaker	HQ
\note	wannju i:nigi gunju-?: daddja la?ambuni
\date	11/Sep/2020

Malyangapa_abbreviations

\abb	**vtr**
\mng	transitive verb
\type	**sub-category**
\nt	transitive verbs are a sub-category of verbs; they take a transitive subject argument in ergative case and a transitive object argument in accusative case.
\cf	**vi, vdi**
\gr	
\date	11/Sep/2020

How do you analyse the corpus?

Typically involves computer software:

1. Praat – used for time-aligned transcription and acoustic analysis of audio
2. ELAN – used for time-aligned transcription of audio and video, translation (word, morpheme, sentence), annotation (part of speech, grammatical function, speaker gender, gesture). Very fast searching over large corpora, can play tokens instantly, files stored as XML so manipulable by other software
3. FLEx – used for aligned morpheme-by-morpheme analysis and glossing of sentences, creation of concordances (e.g. KWIC key word in context), creation of linked lexicon, can export multilingual dictionary in various formats, including HTML (for web) and LIFT (for apps and other software)
4. R – used for quantitative analysis of linguistic and non-linguistic variables based on ELAN or FLEx XML files

Archiving

- An archive is a trusted repository with a collection policy and a commitment to:
  - appraise the value of certain materials
  - preserve selected items
  - make known their existence
  - enable access to them (or their ‘content’)
- Archives have a catalogue that presents metadata (data about the data in the archive), often in a standardized format, some have **finding aids**
- Archives have access management protocols
- Many funders and organisations now require that projects archive their materials

Henke & Berez-Kroeker (2016: 411)

"It is difficult to imagine a contemporary practice of language documentation that does not consider among its top priorities the **digital preservation** of endangered language materials. Nearly all handbooks on documentation contain chapters on it; conferences hold panels on it; funding agencies provide money for it; and even this special issue evinces the **central role of archiving** in endangered language work. In fact, archiving language data now stands as a regular and normal part of the field linguistics workflow (e.g., Thieberger & Berez 2011)." [emphasis added]

Note: there is a free online course about archiving at https://archivingforthefuture.teachable.com/, but it does not cover how to use other people's collections or legacy materials

Important

A website is **NOT** an archive!!!

1. Websites are volatile – they come and go, and rarely have the institutional support like an archive does
2. The files on websites can become obsolete and no longer accessible; archives plan for ‘forward migration’ of file formats
3. Access to websites cannot typically be controlled to the fine degree that archives allow
4. Anyone can put anything on the web – archives involve collection policies, selection, and curation (quality control)

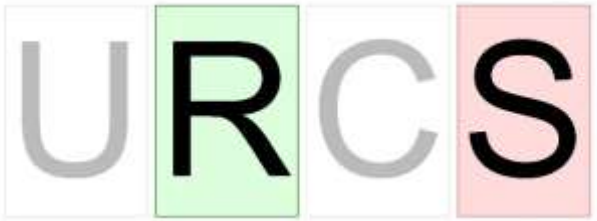
Archive types

1. Classified according to the types of material:
   - **Physical** (analogue) archives – contain paper records, tape recordings, physical objects, e.g. Smithsonian Institution, British Library, Bibliothèque nationale de France
   - **Digital** archives – contain digital files only: audio-visual, text, still images, maps, e.g. ELAR, TLA, AILLA (see DELAMAN for a list)
   - **Mixed** archives – contain analogue and digital materials, e.g. AIATSIS, CLA, ANLA
2. Classified according to scope:
   - **International** – world-wide or multi-country coverage, e.g. ELAR, TLA, BL, BNdeF, AILLA, Pangloss
   - **National** – cover one country, e.g. AIATSIS
   - **Regional** – cover an area in a country, e.g. CLA, ANLA
   - **Local** – cover a town or community, e.g. local museums
   - **Personal** – records of an individual or family

Large international digital – ELAR

Large international digital – TLA DOBES

Archive access management



- Universal – resource available to all, e.g. online
- Register – resource available to registered users
- Closed – resource not generally available (embargoed, “black box”)
- Strict – resource available to users who have been given *individual* access rights for that resource

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

Approaches to language revitalisation

- Often assumed revitalisation = formal language learning (school lessons, immersion, bilingual education)
- Other methods may be easier to establish and to identify resources for:
  - Language landscape
  - Language nests
  - Master-apprentice programmes
  - Cultural immersion and informal learning: camps, seeing and doing
- Many communities are now trying to use descriptive and corpora materials to support language learning – one of the reasons it is important to properly manage your research materials and create a usable corpus that others can access

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 involves dealing with often **complex issues** about the form, content, context, and use of materials and analyses arising from them
- good corpus management principles and practices (e.g. file naming, folder structure, backup, software tools) will make life easier. It is essential to build in archiving plans from the beginning of a project
- 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 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 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

Abbreviations

AIATSIS	Australian Institute of Aboriginal and Torres Strait Islander Studies
AILLA	Archive of the Indigenous Languages of Latin America (UTexas Austin)
ANLA	Alaskan Native Languages Archive
APS	American Philosophical Society
BL	British Library
BNdeF	Bibliothèque nationale de France
CLA	California Languages Archive (UC Berkeley)
DELAMAN	Digital Endangered Languages and Musics Archives Network
ELAR	Endangered Languages Archive (SOAS University of London)
SI	Smithsonian Institution
TLA	The Language Archive (MPI Nijmegen)

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. 2016. Language documentation 20 years on. In Martin Pütz & Luna Filipović (eds.) *Endangered Languages and Languages in Danger: Issues of ecology, policy and human rights*, 147-170. Amsterdam: John Benjamins.

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

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice & Anthony C. Woodbury. 2017. Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics* 56(1), 1-18.

References

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 field materials. *Language Documentation and Description* 21.

Gawne, Lauren, Barbara F. Kelly, Andrea L. Berez-Kroeker & Tyler Heston. 2017. Putting practice into words: The state of data and methods transparency in grammatical descriptions. *Language Documentation & Conservation* 11, 157–189.

Henke, Ryan & Andrea L. Berez-Kroeker. 2016. A Brief History of Archiving in Language Documentation, with an Annotated Bibliography. *Language Documentation & Conservation* 10, 411-457.

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.

McEnnery, Tony & Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory, and Practice*. Cambridge: Cambridge University Press.

References

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)

Rice, Keren. 2006. Let the language tell the story? The role of linguistic theory in writing grammars. In Felix K. Ameka, Alan Charles Dench & Nicholas Evans (eds.) *Catching Language: The Standing Challenge of Grammar Writing*, 235-268. Berlin: Mouton de Gruyter.

Thieberger, Nicholas & Andrea L. Berez. 2011. Linguistic data management. In Nicholas Thieberger (ed.) *The Oxford handbook of linguistic fieldwork*, 90-118. Oxford: Oxford University Press.

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.

References

Woodbury, Anthony C. 2003. Defining documentary linguistics. *Language Documentation & Description* 1, 35-51.

Woodbury, Anthony C. 2011. Language documentation. In Peter K. Austin & Julia Sallabank (eds.), Cambridge handbook of endangered languages, 159–186. Cambridge: Cambridge University Press