
SOAS guest lecture 2022-11-01

Overview of tools and techniques for language documentation, description and revitalisation

Peter K. Austin

Department of Linguistics
SOAS, University of London

© 2022 Peter K. Austin

Creative commons licence

Attribution-NonCommercial-NoDerivs

CC BY-NC-ND

www.peterkaustin.com

Overview

- Research components and tools
 - Planning
 - Recording, selecting
 - Transcribing, translating
 - Annotating (glossing, categorizing)
 - Metadata recording and management
 - Archiving and mobilisation
 - Managing your materials
 - Survey of analysis tools and functions
 - Comparisons
-

Research project organisation

Component	Tool
1. Planning	
2. Recording – audio, video, photos, text, metadata	
3. Selecting, Editing, Filenaming	Audacity, iMovie, Handbreak
4. Adding value	
Transcription	Praat, ELAN
Translation	ELAN, FLEx
Annotation/Glossing	FLEx
Metadata	Excel, Arbil, SayMore

Project organisation (cont.)

Component	Tool
5. Archiving	
6. Mobilisation	Lexique Pro, CuPed, apps

Project management

Two basic principles to manage your research materials:

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. 2022-07-11_LDSS2022_tools.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
-

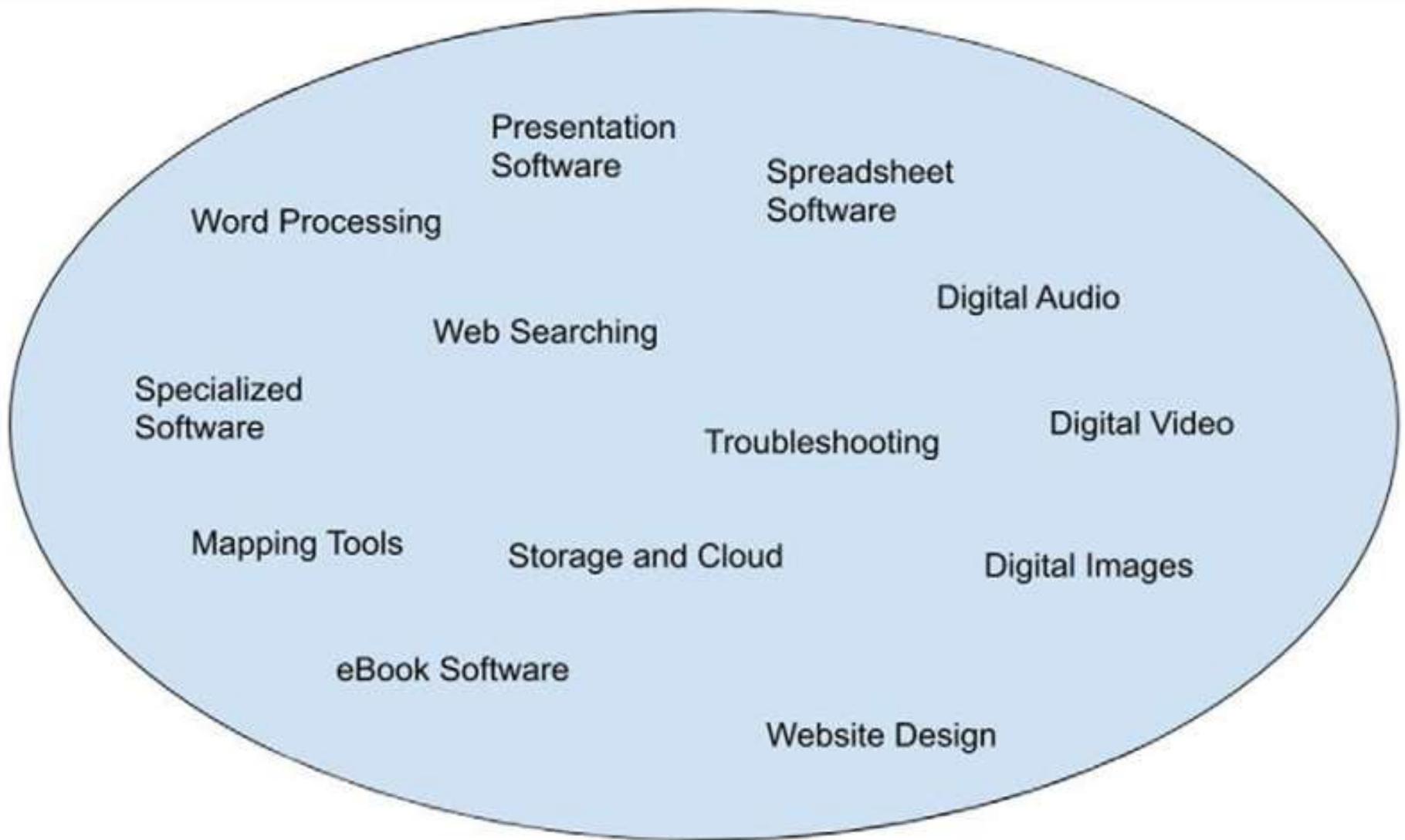


Figure 17.1 The Wide World of Apps. A possible sea of uncharted ‘Apps’ relevant to language revitalization workers. Developing expertise in all areas is daunting, perhaps even an impossible task

After you make an audio or video recording

- You probably need to transcribe it.
 - You may need to translate it.
 - You may want to add other information.

 - Some tools will help you transcribe (“specialist software”)
 - ELAN, Praat, and Transcriber are three that linguists are using these days
 - SayMore also incorporates a transcription component
-

Representation issues with transcription

- Granularity
 - Utterances – NB. **Not** sentences (abstract units of analysis)
 - Conversational turns – what about overlap?
 - Phonological phrases or intonation units (how defined?)
- Types of transcription
 - Segmental
 - Non-segmental, e.g. ToBI
 - Gesture, eye-gaze, body posture
 - Non-linguistic – coughs, pfftt ..., silences (length?), e.g. CA

Representation issues with transcription

- Encoding
 - Phonetic
 - Phonemic (but what if analysis changes?)
 - Suprasegmental
 - Orthographic (benefits? problems?)
 - Multiple linked types of transcription (tiers)
 - Time-alignment – why? potential benefits? replicability?
-

ELAN

- “ELAN (EUDICO Linguistic Annotator) is an annotation tool that allows you to create, edit, visualize and search annotations for video and audio data.”
 - links text annotations with audio and/or video data by time alignment
 - one audio stream, up to four video streams
 - ELAN files can be exported in a variety of formats (including to FLEX for interlinearisation, then reimported)
-

What can ELAN do?

- It can help with transcription and translation
 - It can help with your analysis by presenting your data
 - It can help keep you organised by linking the media and data files together
 - It can help you find things in your data (powerful cross-corpus search)
 - It can help if making a product for community members (text, subtitled video)
-

What can't ELAN do?

- It cannot do your transcription
 - It cannot do your analysis
 - It cannot keep you organised
 - It cannot (by itself) make a viewer for community members
 - It is not (unfortunately) very easy to learn
-

Tiers

Segment	Text	Morphological Analysis
Wioi014_0001	jadi, ore' to a:...	
Wioi014_0002	makawengke, oras makawengke.	
	so, if the one who	will get married, when they're going to
	jadi, kalau yang	mokaweng, diwaktu mokaweng
	jadi, ore' to aa	makawengke, oras makawengke.
	jadi ore' to aa	ma- kaw Ce oras ma- kaw Ce
	so or NR um	AV- marr CPL time AV- marr CPL
	conj conj prt int	pref- v prt n pref v prt

Tiers

- Tiers are where you put your annotations
 - Tiers can contain many kinds of annotations, some of the most obvious are:
 - IPA transcription
 - practical orthographic transcription
 - free translations into languages of wider communication
 - morphemes and gloss
 - gesture annotation
 - grammar notes
 - any other information which seems relevant
-

ELAN – plus and minus

- ✓ Handles most audio and video formats
 - ✓ Powerful for annotating and searching
 - ✓ Good compatibility with FLEX/Toolbox
 - ✓ Good exports for web video etc via CUPED or other tools
 - ✓ Good support and user base
 - ✓ Multi-platform, open-source
 - ✗ Difficult to get started – steep learning curve
 - ✗ No inbuilt tools for interlinearising or lexicon building
 - ✗ *Too* powerful/flexible – temptation to add lots of tiers, gets cluttered and confusing
-

Transcriber

- Transcriber is a tool for assisting the manual annotation of speech signals.
 - It provides a user interface for segmenting long duration speech recordings, transcribing them, and labeling speech turns, topic changes and acoustic conditions.
 - <http://trans.sourceforge.net/en/presentation.php>
-

report

speaker#2
 ● ((Yeah)).

speaker#1
 ● {inhale} He's hilarious. {laugh}

speaker#2
 ● He's great.

speaker#1 + speaker#2
 ● 1: {inhale} He's really a trip.
 2: I know. But it really shows you.

speaker#2
 ● I mean, you know, you really don't have to put up with the Anthony's of the world.

speaker#1
 ● ((I-)) You know what, Ann, it's like, I mean, {exhale}

speaker#1 + speaker#2

⏪ ⏩ ⏮ ⏭ ⏯ ⏸ ⏹ know

Resolution



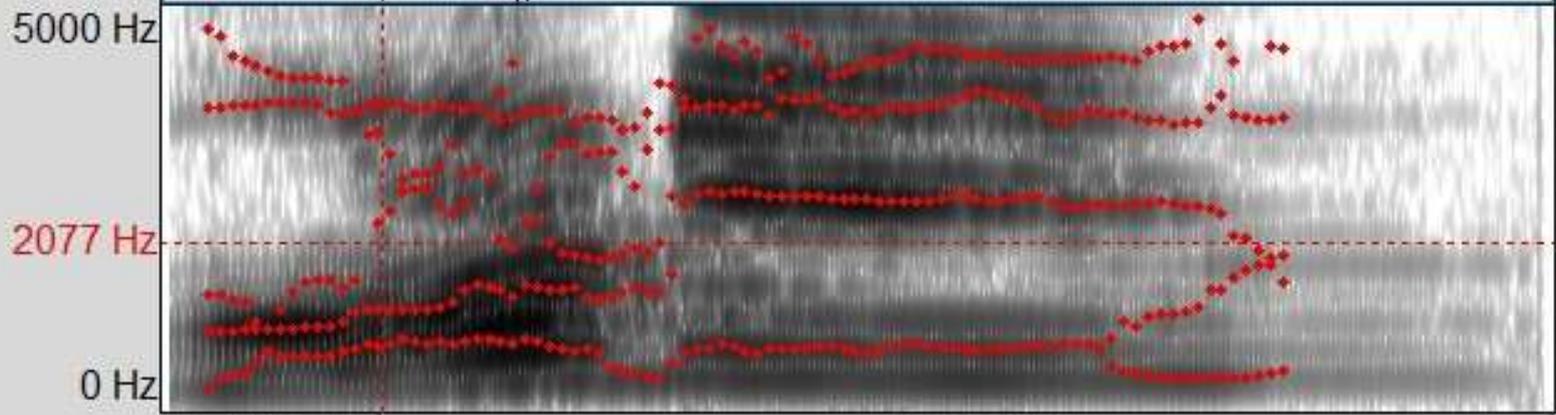
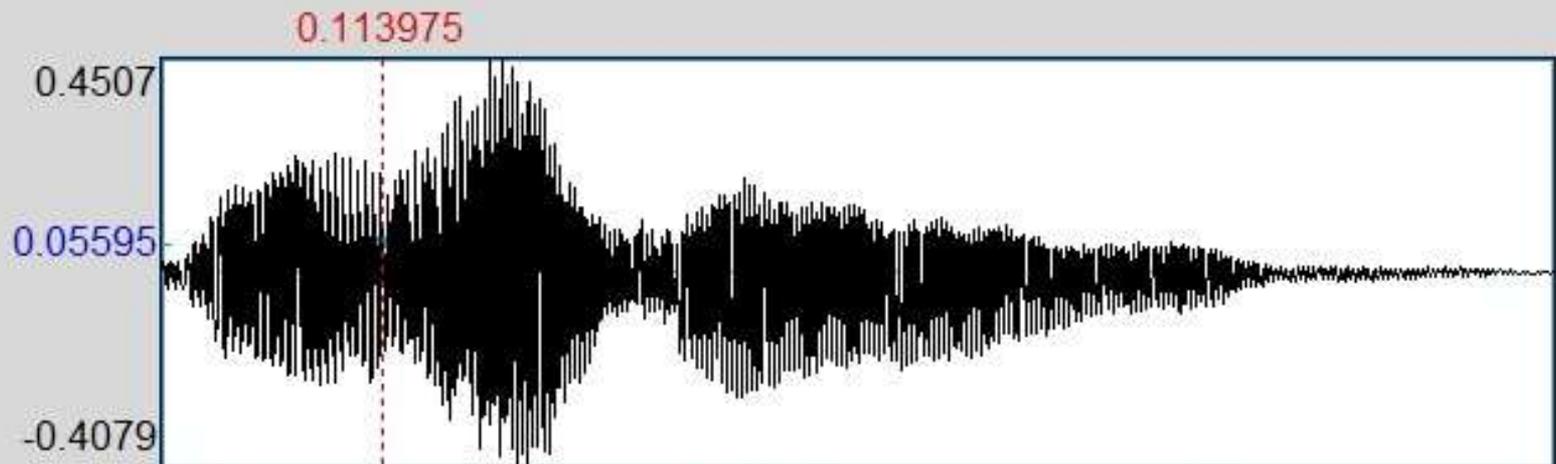
report									
speaker#1	s.	speaker.	speaker#2	speaker#1	speaker#1 +...	speaker#1	.s	speaker#1	spea
{inhale} ...	H	{inhale}.	I mean, you know, you...	((I-)) You know ...	I just didn't know. ...	And the thing is,	{	You know ...	{laugh
... {laugh}	.at.	I know...	... the world.	... mean, {exhale}	I know.	... {laugh}	{	... just-	}

Transcriber plus and minus

- ✓ Relatively easy to set up and use
 - ✓ XML format for easy file exchange
 - ✓ Handles most audio formats
 - ✓ Multi-platform, open source
 - ✗ Lacks video support
 - ✗ Overlapping speech tricky to handle when exporting to FLEX or Toolbox
 - ✗ Not (really) designed for linguists – unlikely to integrate with linguistic analysis tools in the future
-

Praat

- Praat is a free tool for assisting with transcription and acoustic analysis and processing of speech signals.
 - It provides a user interface for listening to speech recordings, studying formants and pitch, transcribing and synthesizing them, and creating pictures
 - Has complex scripts and functions
 - Large supporting responsive community
 - <http://www.fon.hum.uva.nl/praat/>
-



0.113975 0.599993

0 Visible part 0.713968 seconds 0.713968

Total duration 0.713968 seconds

You've transcribed. Now what?

- Grammar analysis
- Lexicon building
- Cultural/ethnographic notes
- ???

A tool that help you do some of these things:

- Fieldworks Language Explorer (FLEX)
– from SIL International
-

Fieldworks Language Explorer

- “FieldWorks is a set of software tools that help manage cultural and linguistic data from initial collection through submission for publication”
 - It can be used to record lexical information and develop dictionaries.
 - It can interlinearize text.
 - The morphological parser provides the user with a way to check the grammatical rules they have recorded against real language data.
 - The grammar information can also be compiled in an automatically generated grammar sketch.
-

Lexicon

-  Lexicon Edit
-  Browse
-  Dictionary
-  Categorized Entries
-  Classified Dictionary
-  Bulk Edit Entries
-  Bulk Edit Senses
-  Reversal Index
-  Bulk Edit Reversals

 Lexicon

 Texts

 Words

 Grammar


Entries

Headword 	Gloss 
Show All 	Show 
-	dividir
gaya	vaidac ser vai
goa	vale
goce	farelo
godama	joelha
godamira	joelha
gogodo	osso
gogoma	ajoelh
gombe	marge
gona	dormi
gopa	ter me
gubudza	sacudi
gula	compr
..	tranda

Entry

 Show Hidden Fields

godamira *V.* Por joelhar a favor de algém Eng kneel for someone

Lexeme Form

godamir

Morph Type

root

Citation Form

godamira

Entry Type

Main Entry

Sense 1

Gloss

 Por **joelhar**
 Eng **kneel**

Definition

 Por joelhar a favor de algém
 Eng kneel for someone
Grammatical Info. Verbo 

Example

Translation

Reference

Semantic Domains

Lexical Relations

Parsing Note

Kalaba - Ls-hovland - FieldWorks Language Explorer

File Edit View Data Insert Format Tools Parser Window Help

English

Texts **Texts** **Text**

Edit
Document

My Green Mat

Lexicon

Texts

Words

Gram...

Lists

My Green Mat

This is a simple sentence about seeing a green mat.

Word	pus	yalola		nihimbilira			
Morphemes	pus	yalo	-la	ni-	him-	*bili	-ra
Lex. Entries	pus ₁	yalo	-la	ni-	hiN-	*bili	-ra
Lex. Gloss	green	mat	1SgPoss	1SgSubj	3SgObj	to.see	Pres
Lex. Gram. Info.	adj	N (1)	N:(Possessor)	V:(Subject)	V:Object	trans (1)	sta:Tense
Word Gloss	green	mat		I see			
Word Cat.	mod	N		V			

Free: I see my green mat.

Queue: [-/-] No Parser Loaded

1/1

FLEX plus and minus

- ✓ Solid data structure using XML
 - ✓ Very powerful parsing and grammatical analysis tools
 - ✓ Designed to hold all your linguistic and cultural data and notes
 - ✗ Poor handling of media
 - ✗ Large application, memory hog
 - ✗ Windows only
-

Another dictionary tool – We Say

- WeSay helps non-linguists build a dictionary in their own language.
 - It has various ways to help native speakers to think of words in their language and enter some basic data about them (no backslash codes, just forms to fill in).
 - Designed for teamwork – one ‘advanced’ user does the complicated set-up work, very simple interface for other users
-



Home

Dictionary Browse & Edit

Actions

Collect Words By Semantic Domain

bth



abit ab'it water carrying basket

- abit
- abo-abo
- aboh
- abur
- abus
- abūs
- abut
- abūt
- abūt kupong
- adat
- addi
- adis
- adoh
- adoi
- adu
- adu-adu
- adud
- adūd

Word

bth	abit	
ipa	ab'it	

Meaning 1

en	water carrying basket	
----	-----------------------	--

Picture



[Pictures & Images...](#)

POS

noun

Example

bth

Meaning 2

en

New Word

Delete This Word

Show Uncommon Fields

We Say plus and minus

- ✓ Very simple to use
 - ✓ Will run on netbooks and other low-powered machines
 - ✓ Good data structure
 - ✓ Easy export via Lexique Pro for print/web
 - ✗ No tools for interlinearising or analysis
 - ✗ Limited media support
 - ✗ Windows only
-

Comparison of programs

	Transcriber	ELAN	Toolbox	FLEx	WeSay
Audio time-alignment	✓	✓	✗	✗	✗
Video time-alignment	✗	✓	✗	✗	✗
Multi-tier annotation	✗	✓	✓	✓	✗
Interlinear support	✗	✗	✓	✓	✗
Lexicography	✗	✗	✓	✓	✓✗
Word collection	✗	✗	✓	✓	✓
Simple to learn	✓	✗	✗	✓✗	✓
Special char. input	✗	✓	✓	✓	✓
XML data	✓	✓	✗	✓	✓

Managing metadata

- There are a few specialist programs that can be used to manage metadata (many researchers use spreadsheets)
 - Arbil (from MPI Nijmegen) can be used online or stand alone for IMDI metadata
 - SayMore (from SIL) can be used to harvest metadata from files and then say more about it
-

SayMore

EdoloSample - SayMore 2.1.144 (Beta)

Project Session Person Progress Help

Sessions People Progress

Sessions

Id	Title	Stages	Status
ETR009	The story behind how we catch fish with poison bark	██████	○

Open ▾ Rename... Convert...

Name	Type	Date Modified	Size	Duration
ETR009.session	Session	2/12/2013 9:07:1...	805 B	
ETR009_Careful...	Audio	12/20/2012 11:47...	302 KB	00:00:25
ETR009_Tiny.m...	Video	12/20/2012 11:47...	389 KB	00:00:10
SceneAroundC...	Image	12/20/2012 11:47...	81 KB	
SceneHouse.JPG	Image	12/20/2012 11:47...	87 KB	

▾ Add Files...

Session

Status & Stages Notes

Id ETR009	Date 6/ 6/2010 ▾
Title The story behind how we catch fish with poison bark	Setting Sitting on their fron porch
People Awi Heole, Ilawi Amosa ▾	Location Huya
Genre narrative ▾	Access Open
Situation Walking with Ilawi, we passed the plant they use to make poison for fishing, and I asked him if he'd tell me about how to do it. We went to his house, where Awi joined us, and the two of them took turns telling parts of the "whole story".	Abstract
Custom Fields	
Field	Value

New New From Device... New From Recording...

9:07 PM 2/12/2013

Mobilisation tools

- Lexique Pro – tool for making printed and online (web) dictionaries
 - Reads FLEx files and converts them to Word documents (double column, headers, pictures) or HTML files (basic interface) or standalone Windows executable
 - Creates reverse finder list and semantic fields thesaurus
 - Interface in multiple languages
-

The End
