LDSS2022 11th July 2022

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

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

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

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.
- ELAN, Praat, and Transcriber are three that linguists are using these days
- SayMore also incorporates a transcription component

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
- 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


00:00:01.000
00:00:02.000
00:00:03.000
00:00:04.000
00:00:05.000
ref@BH [11]
t@BH [11]
fe@BH [11]
fn@BH [11]
w@BH [33]
m@BH [49]
g@BH [49]
p@BH [49]
Wioi014_0001
Wioi014_0002
jadi, ore' to a:...
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

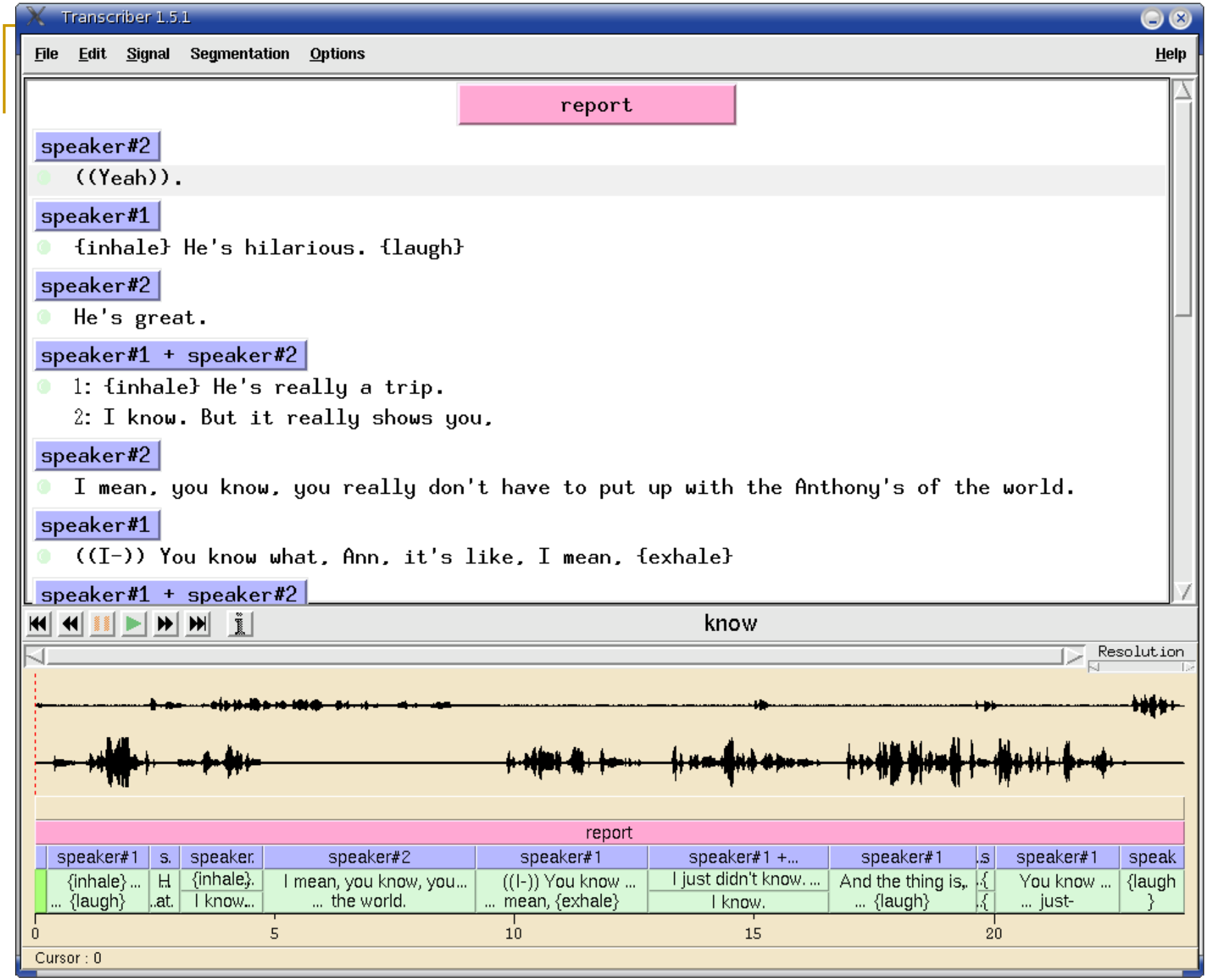
ELAN – plus and minus

- 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


Transcriber 1.5.1
File Edit Signal Segmentation Options Help
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 speak
{inhale} ... {laugh}
H. at.
{inhale}. I know...
I mean, you know, you... ... the world.
((I-)) You know ... ... mean, {exhale}
I just didn't know. ... I know.
And the thing is,. ... {laugh}
.{ .{
You know ... ... just-
{laugh }
0 5 10 15 20
Cursor : 0

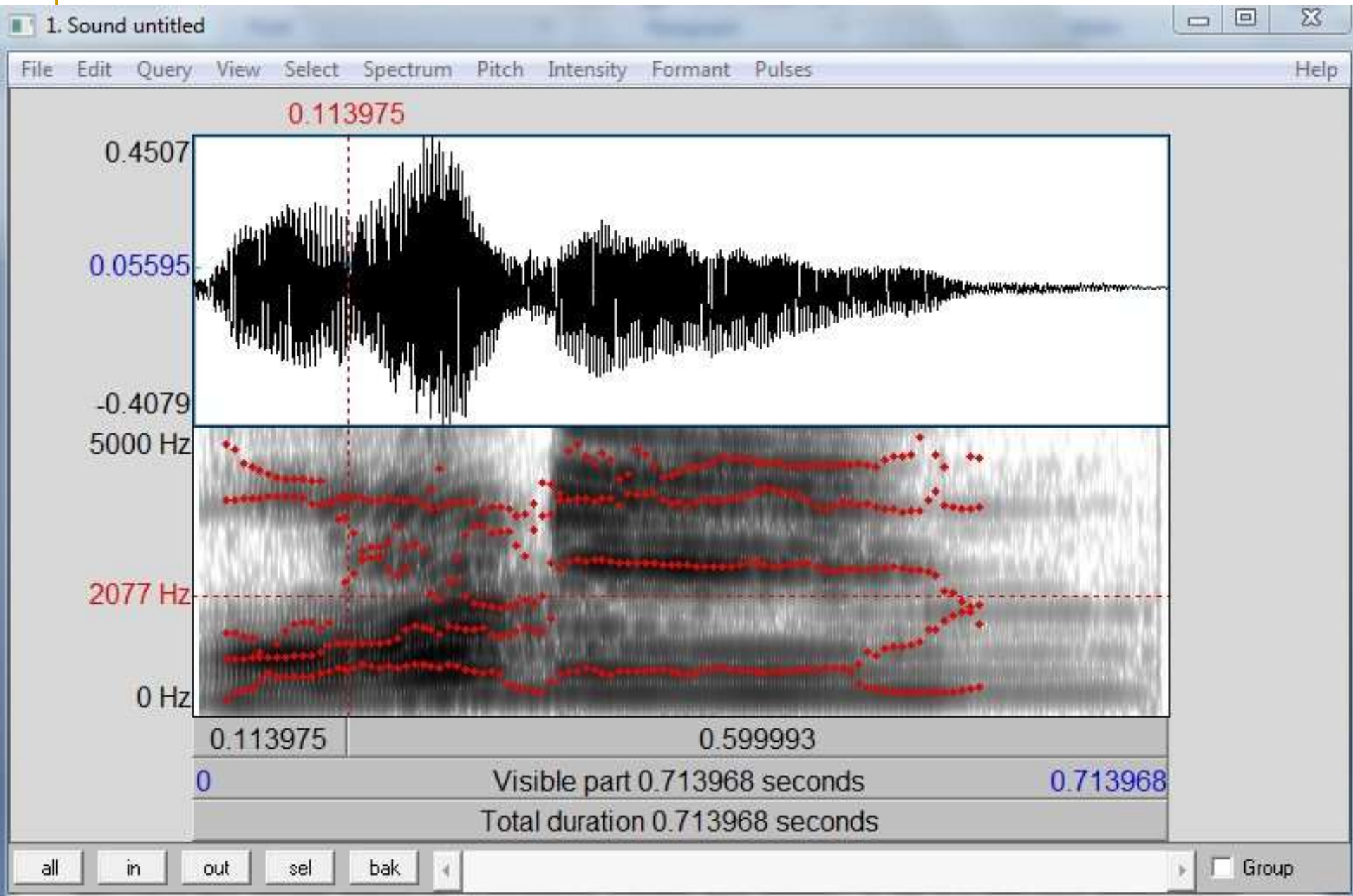
Transcriber plus and minus

- 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/


1. Sound untitled
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses Help
0.113975
0.4507
0.05595
-0.4079
5000 Hz
2077 Hz
0 Hz
0.113975
0.599993
0
Visible part 0.713968 seconds
0.713968
Total duration 0.713968 seconds
all
in
out
sel
bak
Group

You've transcribed. Now what?

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

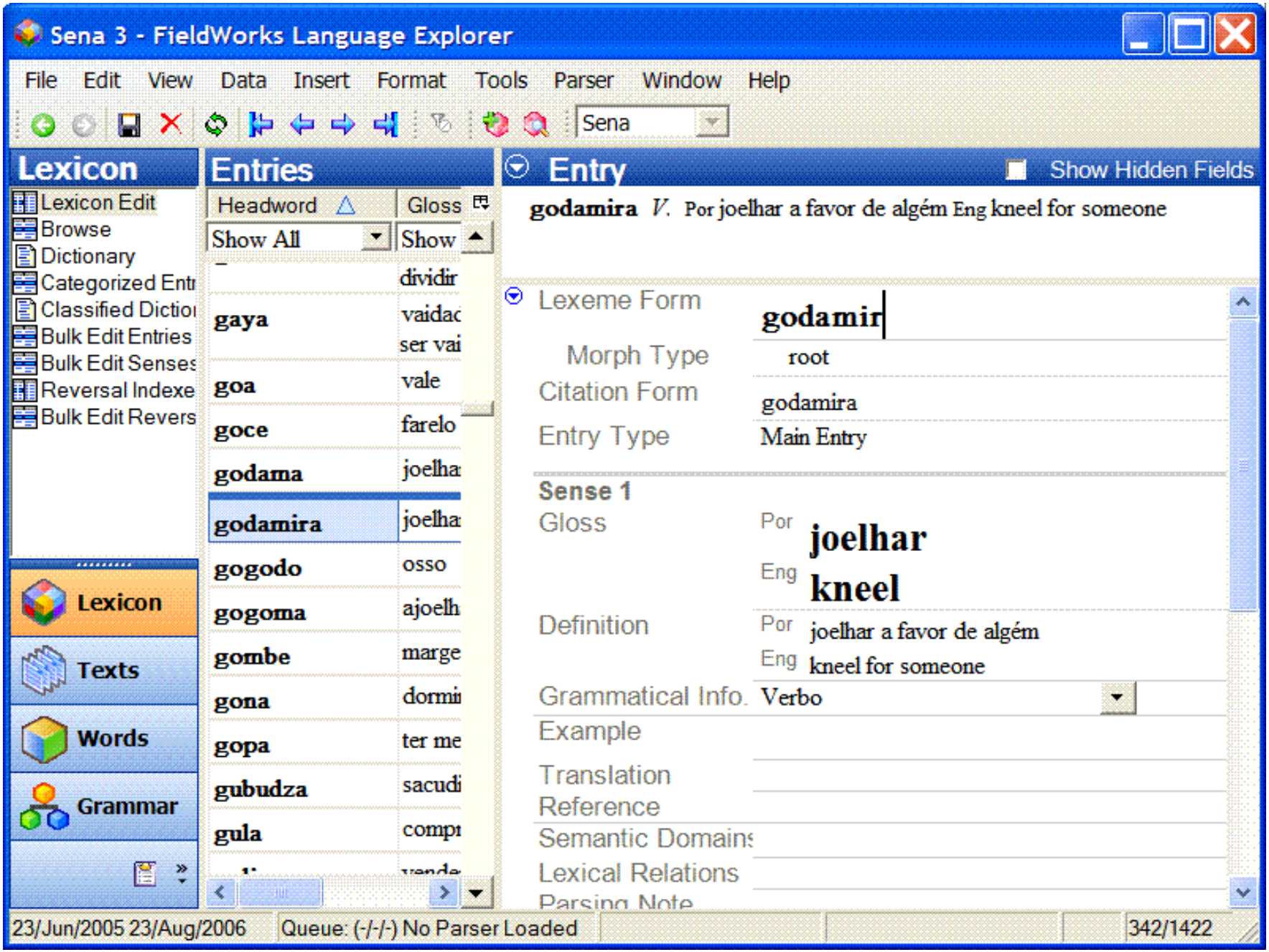
A tool that help you do some of these things:

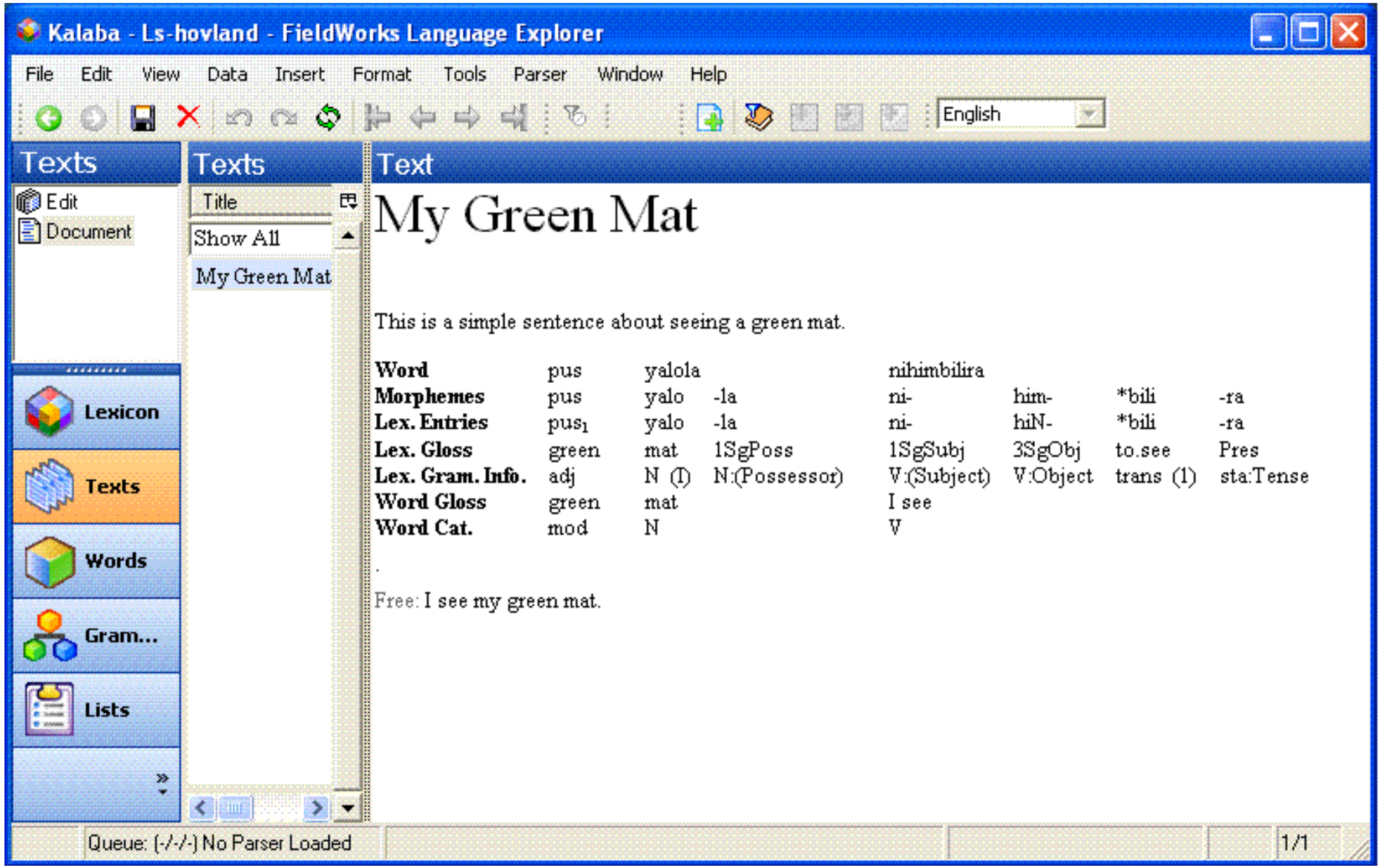
- Fieldworks Language Explorer (FLEX)

– 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.


Sena 3 - FieldWorks Language Explorer
File Edit View Data Insert Format Tools Parser Window Help
Sena
Lexicon
Lexicon Edit
Browse
Dictionary
Categorized Entr
Classified Dictio
Bulk Edit Entries
Bulk Edit Senses
Reversal Indexe
Bulk Edit Revers
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
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
23/Jun/2005 23/Aug/2006
Queue: (-/-/-) No Parser Loaded
342/1422


Kalaba - Ls-hovland - FieldWorks Language Explorer
File Edit View Data Insert Format Tools Parser Window Help
English
Texts
Edit
Document
Lexicon
Texts
Words
Gram...
Lists
Texts
Title
Show All
My Green Mat
Text
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 pus1 yalo -la ni- hiN- *bili -ra
Lex. Gloss green mat 1SgPoss 1SgSubj 3SgObj to.see Pres
Lex. Gram. Info. adj N (I) 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

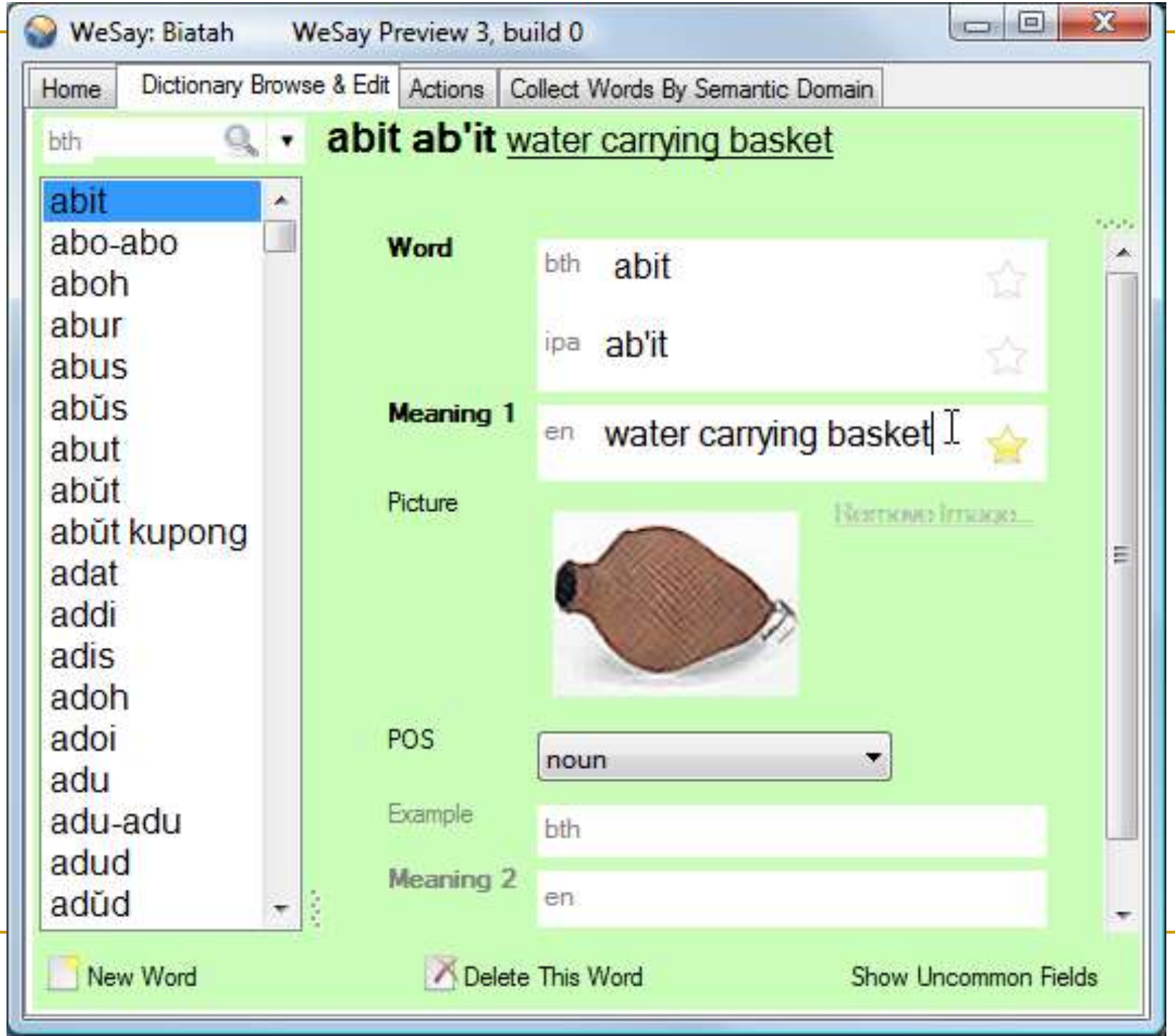
FLEX plus and minus

- 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


WeSay: Biatah
WeSay Preview 3, build 0
Home
Dictionary Browse & Edit
Actions
Collect Words By Semantic Domain
bth
abit
abo-abo
aboh
abur
abus
abŭs
abut
abŭt
abŭt kupong
adat
addi
adis
adoh
adoi
adu
adu-adu
adud
adŭd
abit ab'it water carrying basket
Word
bth abit
ipa ab'it
Meaning 1
en water carrying basket
Picture
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

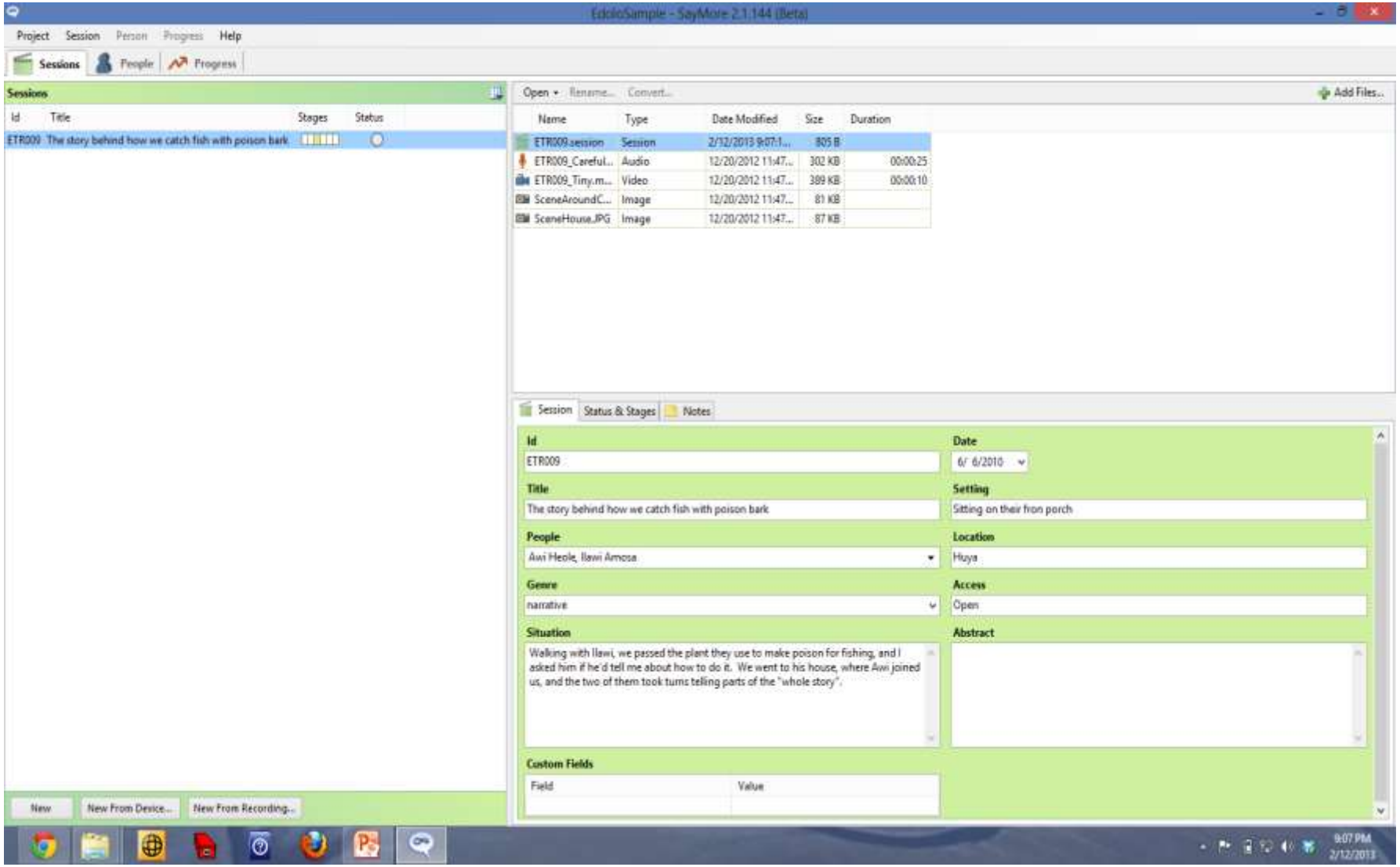
Comparison of programs

	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
New
New From Device...
New From Recording...
Open
Rename...
Convert...
Add Files...
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
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
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

More on ELAN

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
  - sociolinguistic information
  - grammar notes
  - any other information which seems relevant

Tiers and types

- Every annotation tier must be assigned a **Linguistic Type** which tells Elan what type of information the tier contains.

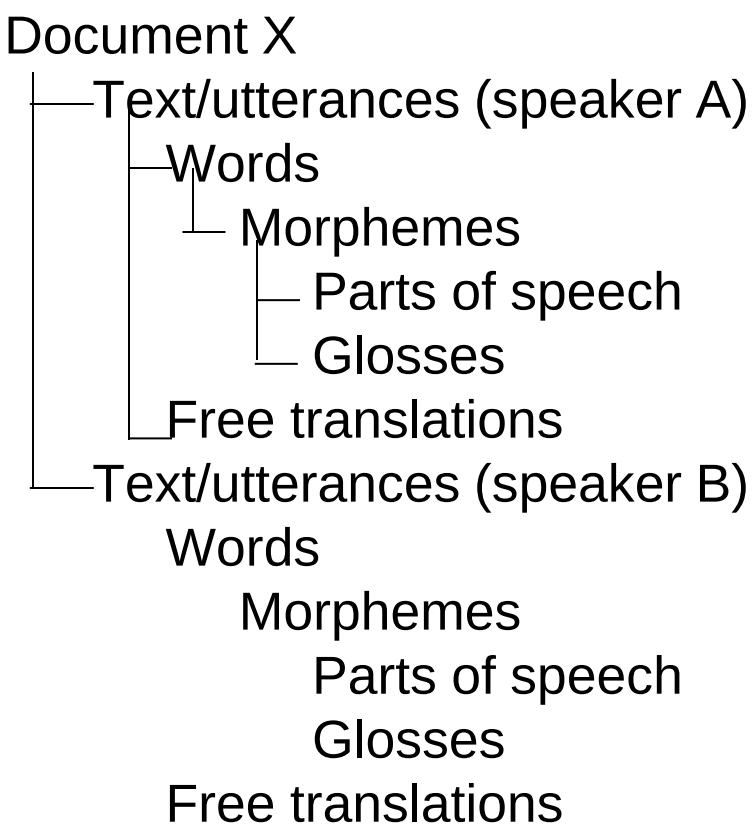
  (It would be better if these were called ‘tier types’).

- You can set up and name the Linguistic Types — e.g. reference number, utterance, word, PoS, translation, and so forth.
- They can have **Controlled Vocabularies** (eg. select a PoS from a drop-down menu)
- The Linguistic Types are defined by **Stereotypes** which basically specify the way that tiers relate to the time line, and to each other
- The Stereotypes (it would be better to call them ‘Tier Categories’ or something) are basically that a tier will be independent, or hierarchically associated with other tiers

Stereotypes

- **None**: The annotation on the tier is linked directly to the time axis (eg. intonation units/sentences - a transcription or a reference number).
- **Time Subdivision**: The annotation on the parent tier can be subdivided into smaller units, which, in turn, can be linked to time intervals (eg. words). There cannot be gaps between units.
- **Symbolic subdivision**: Similar to Time Subdivision, except that the smaller units cannot be linked to a time interval (eg. morphemes within words).
- **Included In**: like Time Subdivision but there can be gaps (eg. words, with silence between them).
- **Symbolic association**: one-to-one association with a parent tier, eg. transcription with ref field, gloss and morpheme, free translation with sentence.

Tier dependencies – parents and children



Stereotypes:
(none)
(time subdivision)
(symbolic subdivision)
(symbolic association)
(symbolic association)
(symbolic association)

Creating Linguistic Types

1. File Menu > New

(Locate a sound/video file. You should see a wave/video file display)

2. Type Menu > Add New linguistic type

3. Type Name: **text**, Stereotype: none > Add
(this will be for sentences)

4. Type Name: **words**, Stereotype: Time Subdivision > Add
(this will be for words)

5. Close

Creating Tiers

- Create a tier for transcription
  – Tier > Add New Tier > type the name **text@A** (this will be for Speaker A)
  (make sure Linguistic Type is text)
  > Add
- Create a tier for words
  – (while Add Tier box is still open) type the name **words@A**
  (make sure Parent Tier is text@A and Linguistic Type is words)

  > Add.
- Close.

Entering and editing annotations

Listen to the sound file (and look at the waveform) and think about breaking the speech up into units (sentences, intonation units). There are two ways to do this:

1. Straight into the tier
   - Select a time span containing a unit: click, hold and drag with the mouse > play the selection (possibly toggle the Loop Mode).
   - Right click > New Annotation here (for the active tier) > transcribe the unit > Enter (to save) or Esc (not to save).
2. Segmentation mode.
   - Edit > Segmentation. A new window appears.
   - Play the sound while pressing Enter to segment (similar to Transcriber).
   - Click Apply to save the segmentation.

- Changing boundaries:
  - click an annotation to select it > draw the new boundary > Enter (to save)
  - select an annotation > press & hold Alt and drag the boundary of the active annotation to a new place.

Now do the same for Speaker B

- Create a tier for transcription
  – Tier > Add New Tier > type the name **text@B** (this will be for Speaker B)
  (make sure Linguistic Type is text)
  > Add
- Create a tier for words
  – (while Add Tier box is still open) type the name **words@B**
  (make sure Parent Tier is text@B and Linguistic Type is words)

  > Add.
- Close.

Adding free translation

- first add a new type **translation**
  - Type > Add New linguistic type
  - Type Name: translation, Stereotype: Symbolic Association > Add
- now add tier translation
  - Tier > Add New Tier > type the name **translation@A** (this will be for Speaker A)
  - (make sure Linguistic Type is f)
  - Add
- The tier should be there. Double click in it to enter sentence translations corresponding to the segments in text. Don’t forget Enter!
- Now do the same for **translation@B**
- Now you can transcribe and translate all the recording