ELAN Linguistic Annotator

Introduction to Speech and Video Annotation with ELAN

http://www.lat-mpi.eu/tools/elan

Introduction to ELAN

- written in Java programming language
- stores transcriptions in XML format (.eaf)
- available for Windows, Mac OS X, Linux
- sources available
- current version 4.4.0

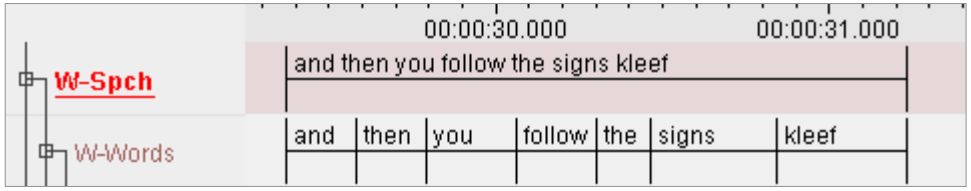
ELAN - Key concepts

Tiers, linguistic types, annotations, controlled vocabularies

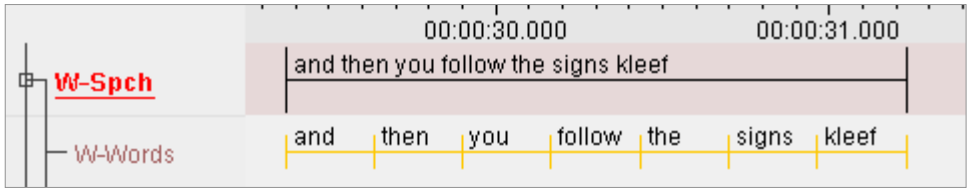
- a tier logically groups annotations that share the same constraints on structure, content and / or time alignment characteristics
- tiers can be part of a tier hierarchy, a tier can have a parent tier and dependent tiers
- an annotation document can contain any number of tiers and annotations
- a controlled vocabulary is a list of predefined annotation values on a tier

ELAN - Key concepts

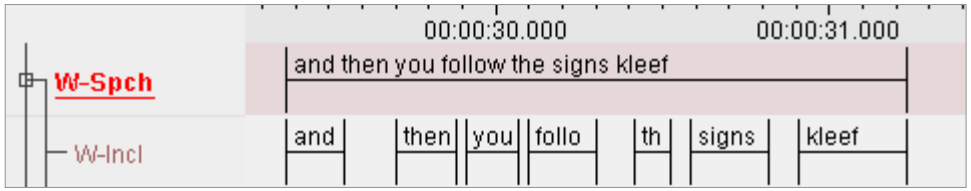
Inter-tier relations



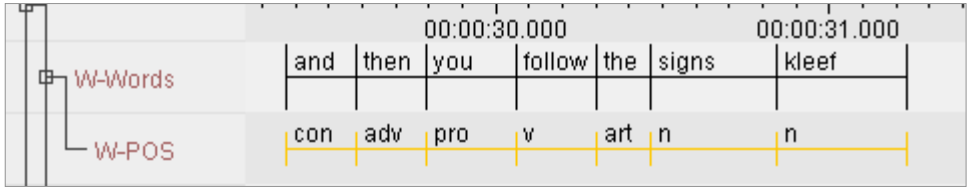
- "Time Subdivision"
  Time aligned subdivision of parent interval, no gaps



- "Symbolic Subdivision"
  Ordered, symbolic subdivision of parent interval



- "Included In"
  Time aligned subdivision of parent interval, gaps are allowed

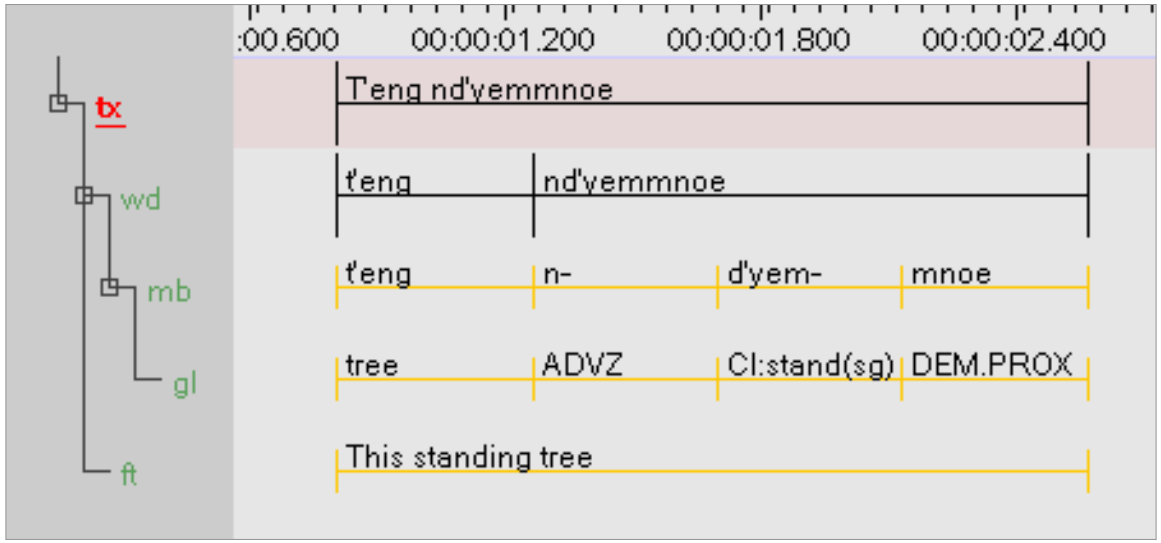


- "Symbolic Association"
  One-to-one association with the parent annotation

Example transcription

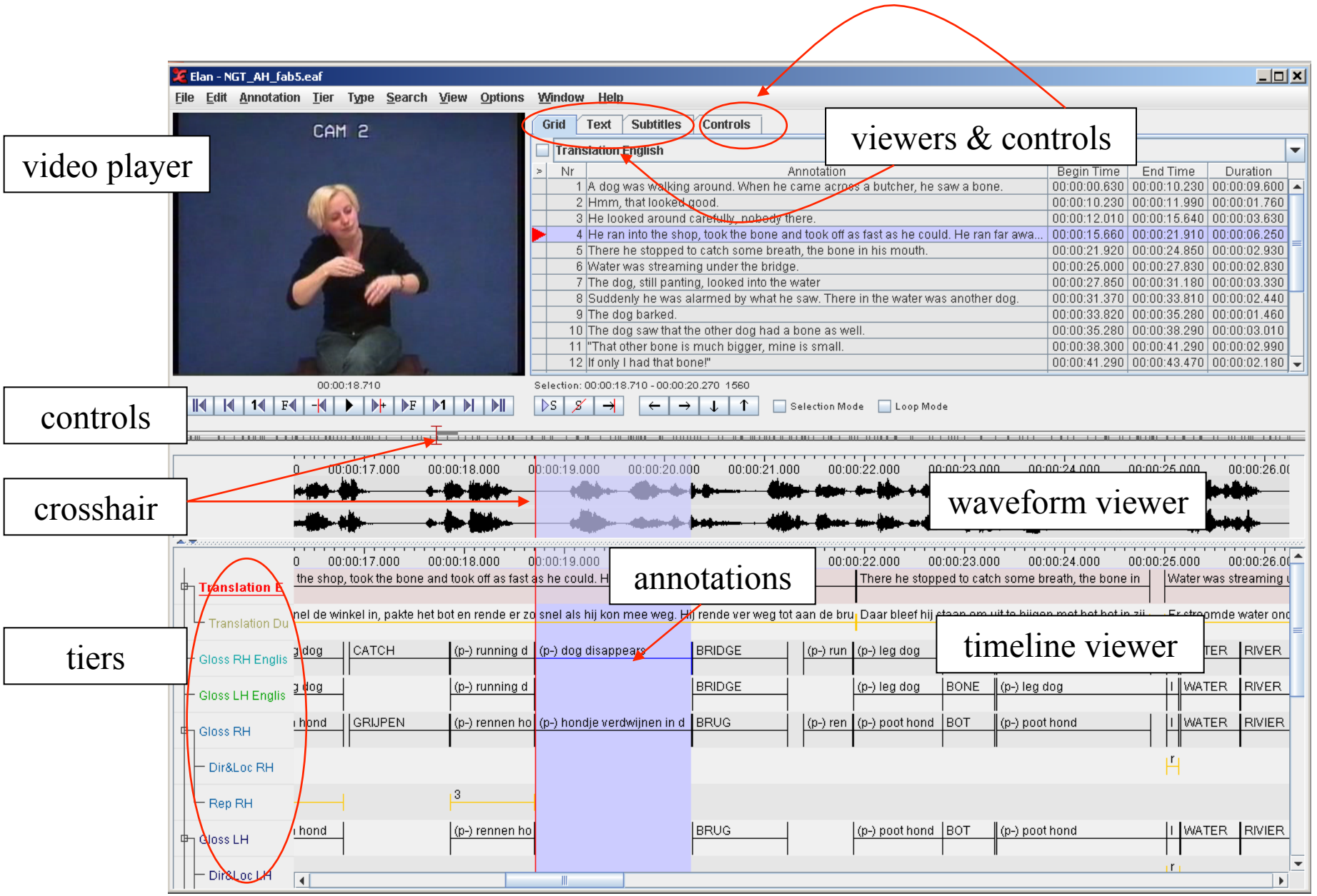
tx (text)	*T'eng nd'yemmnoe.*			
wd (word)	*t'eng*	*nd'yemmnoe*		
mb (morpheme break)	*t'eng*	*n-*	*d'yem-*	*mnoe*
gl (gloss)	*tree*	*ADVZ*	*Cl:stand(sg)*	*DEM.PROX*
ft (free translation)	*This standing tree.*			

interlinear text example (text editor, Shoebox)

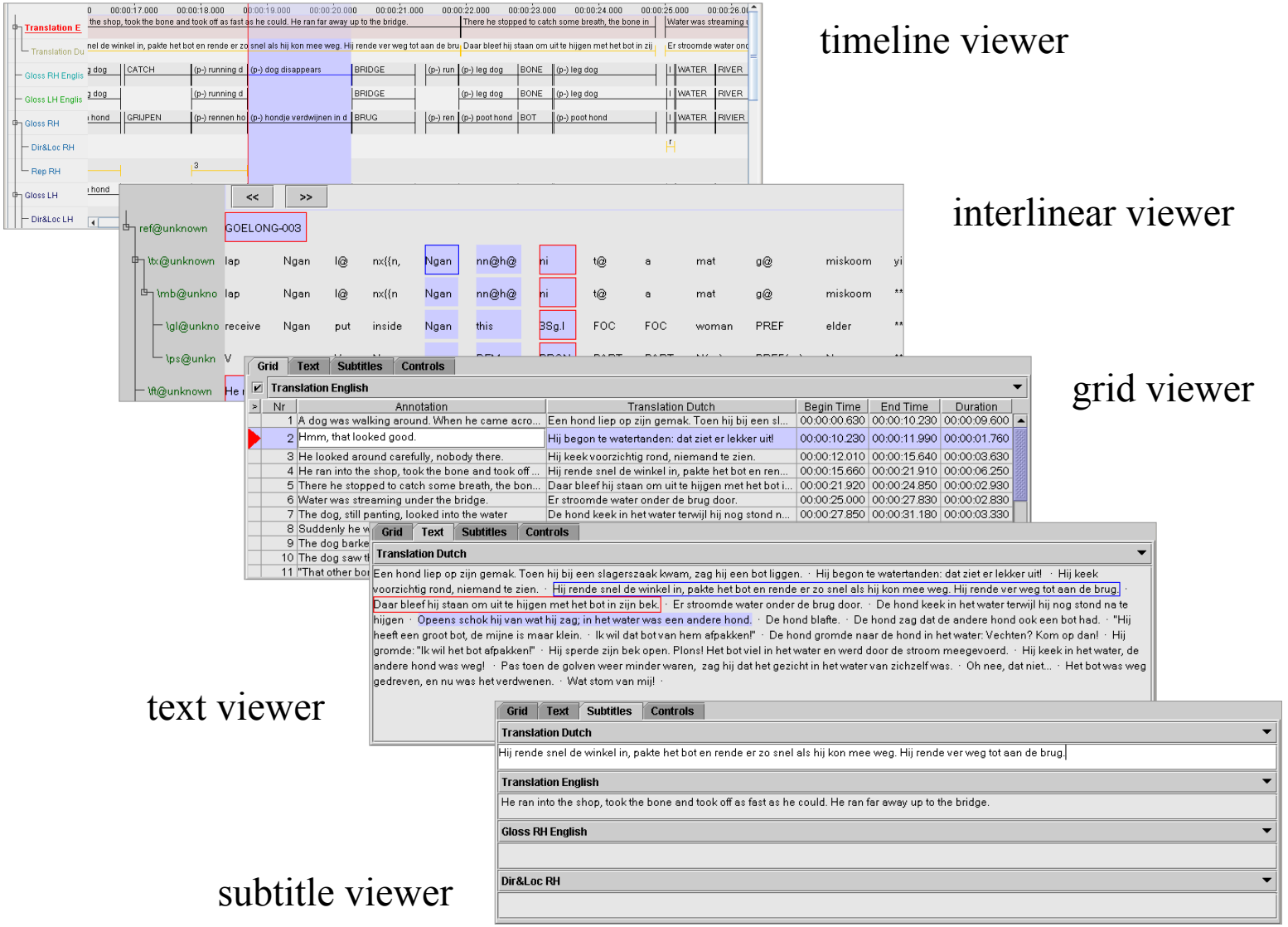


same example in ELAN

ELAN - Main window

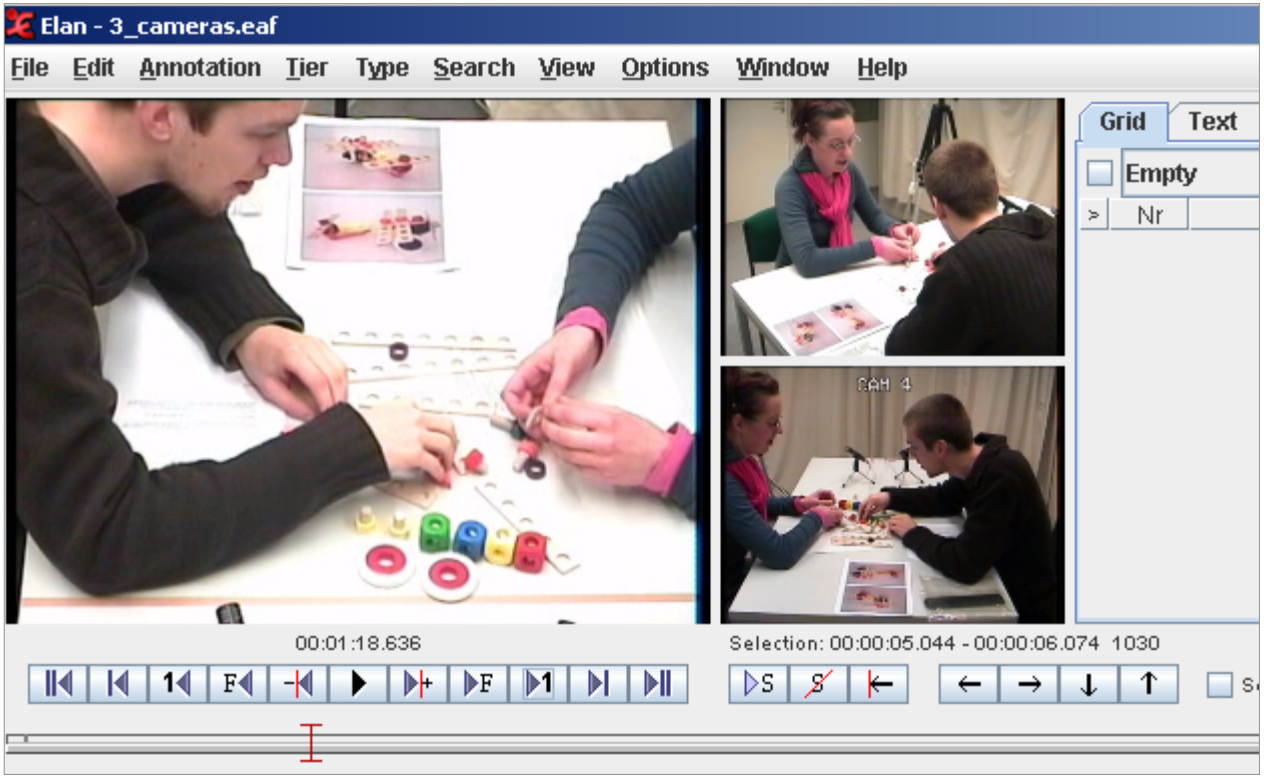


Multiple synchronized viewers



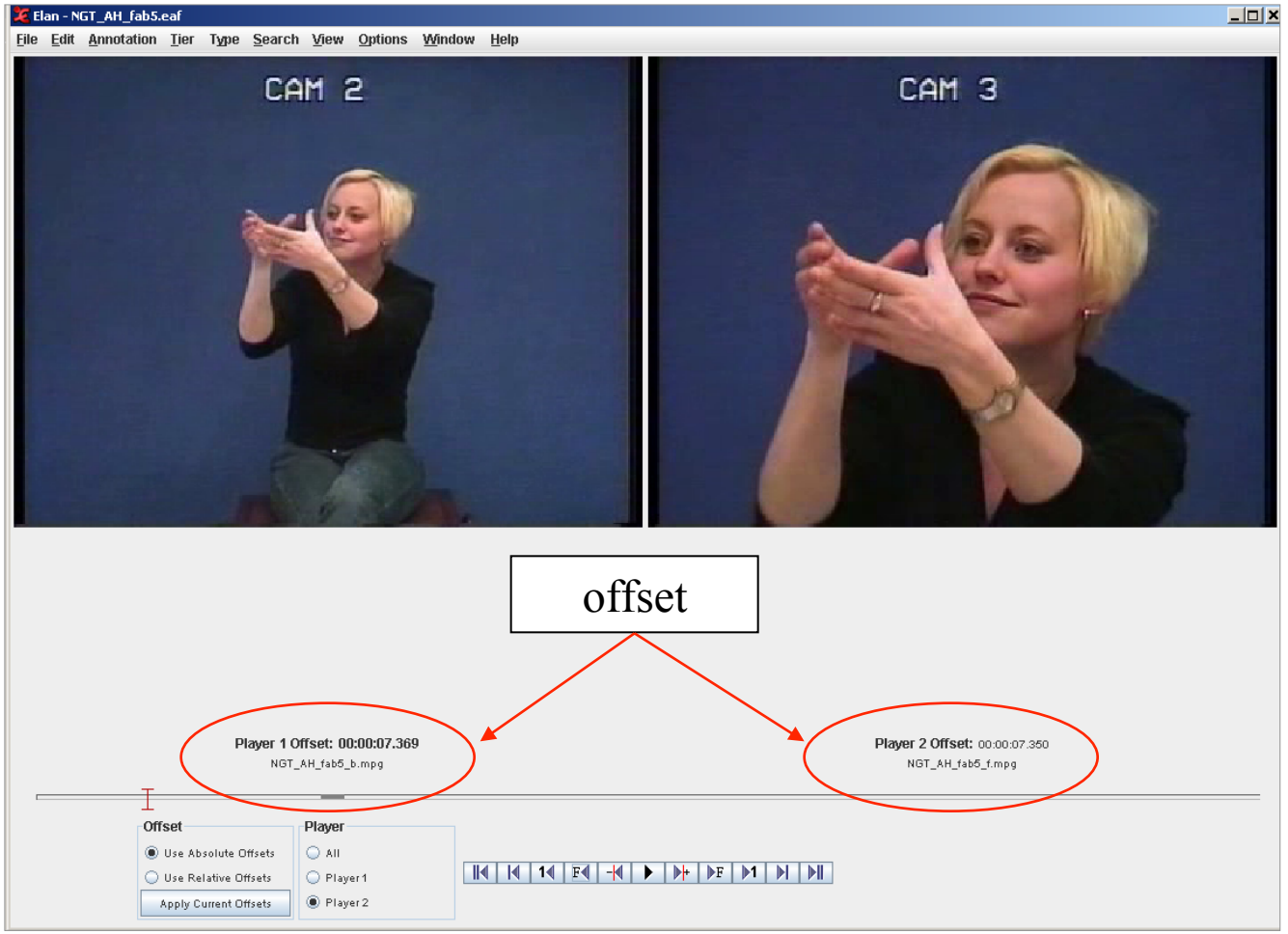
Multiple video support

- support for up to 4 videos
- videos are detachable



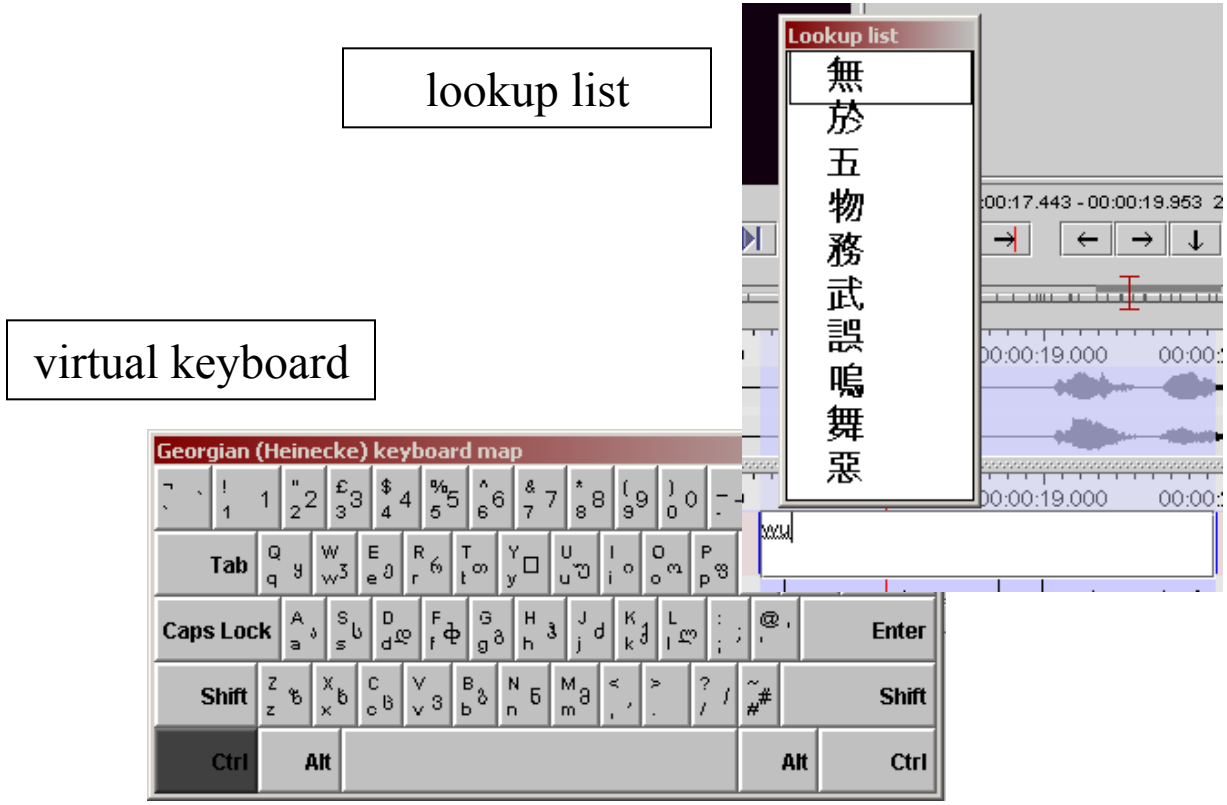
Synchronization of multiple videos

- determine an offset per video



Unicode support

- input and rendering of a wide range of script systems



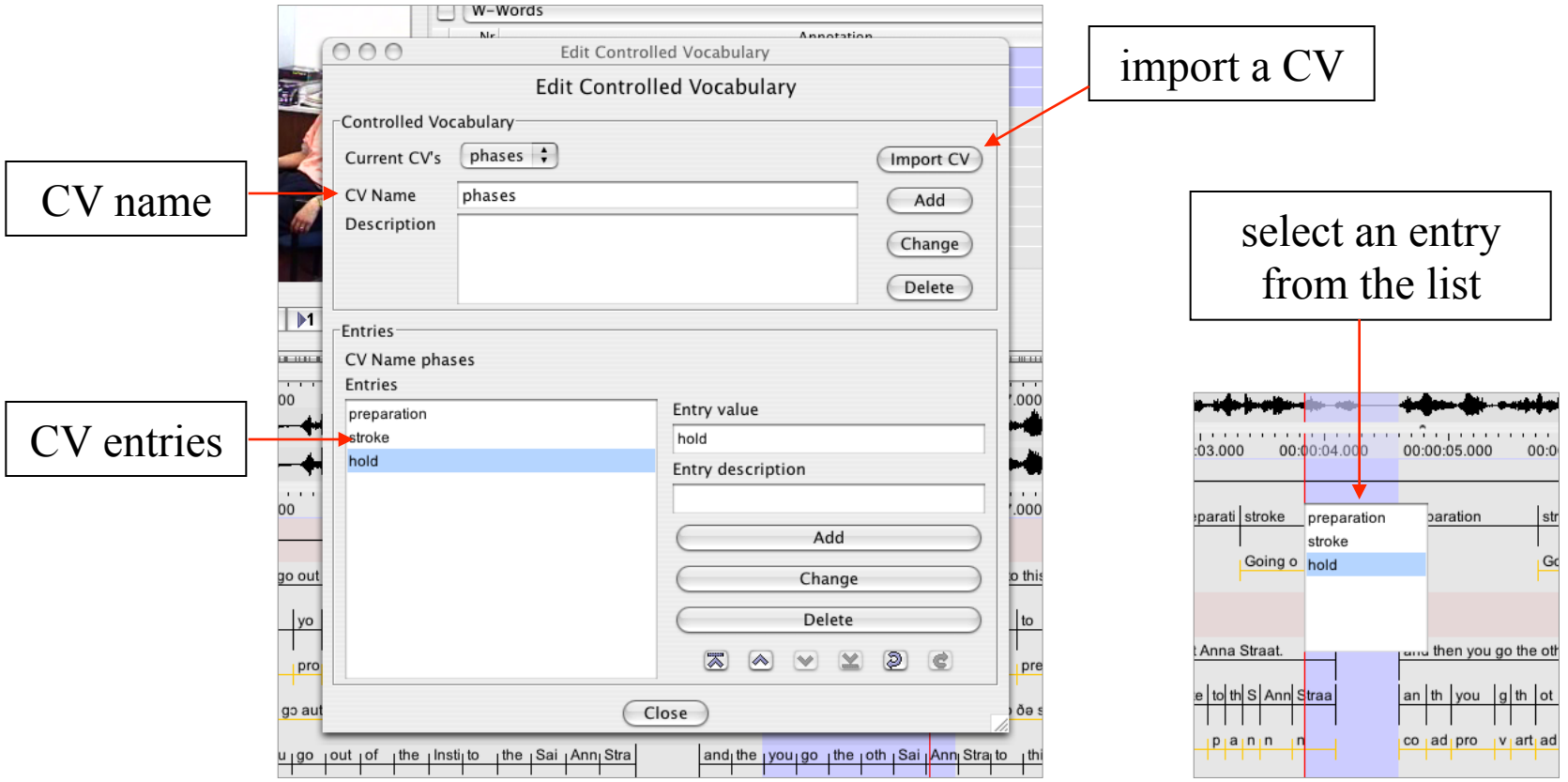
Alternatives:

Windows: Keyman (http://www.tavultesoft.com/)

Mac OS X: Ukelele (http://scripts.sil.org/cms/scripts/)

Use of controlled vocabularies

- user definable, lists of possible annotation values for a tier, e.g. a list of gesture phases (preparation, stroke, hold etc.)

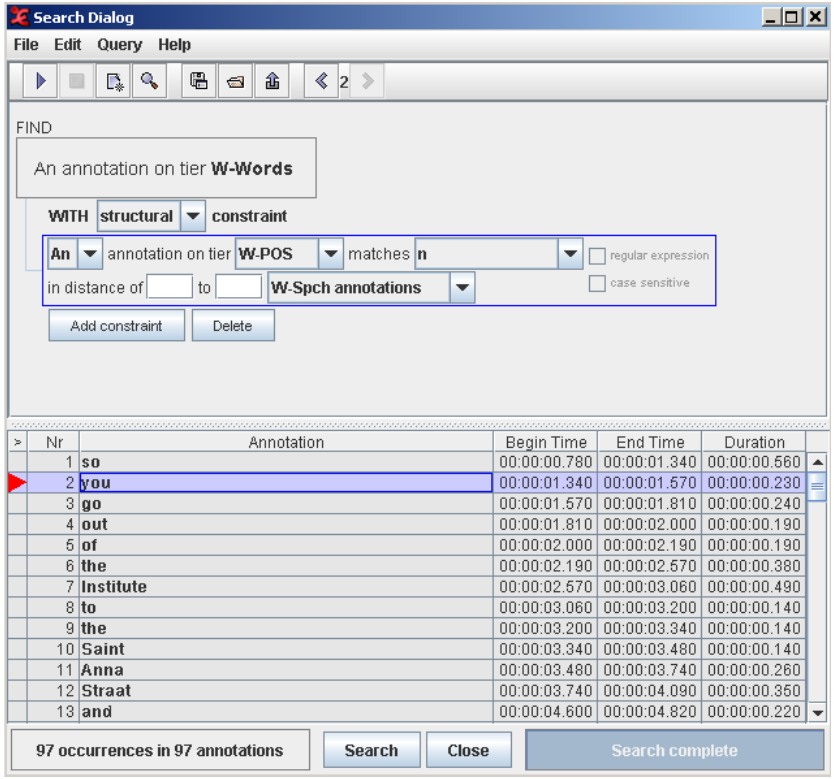


ISO DCR support

- ISO Data Category Registry
- Give annotation concepts a universal tag:
  - Controlled vocubulary, e.g. POS
  - Tier type, e.g. morphological analysis
  - Individual annotation, e.g. metaphore
  - …
- Under development but already usable

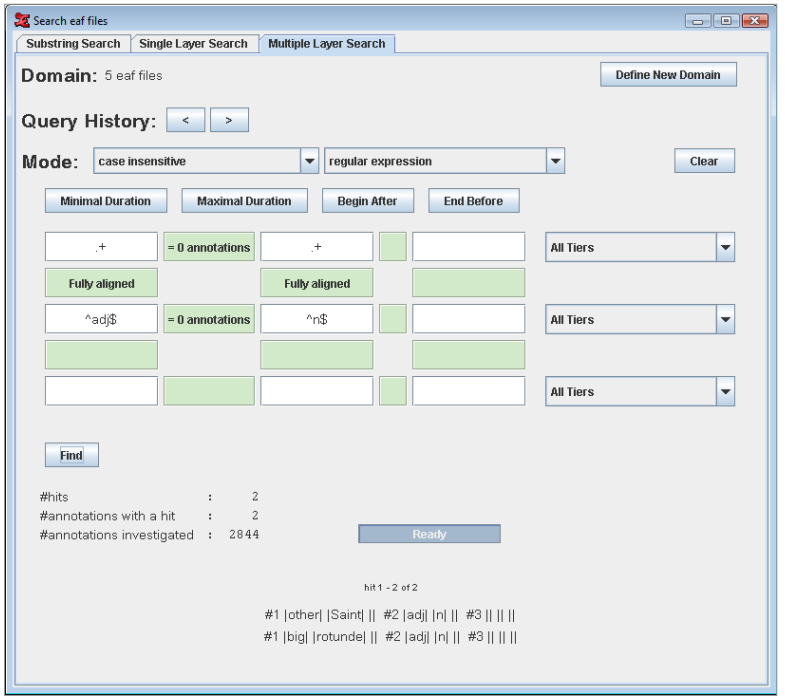
Search in ELAN

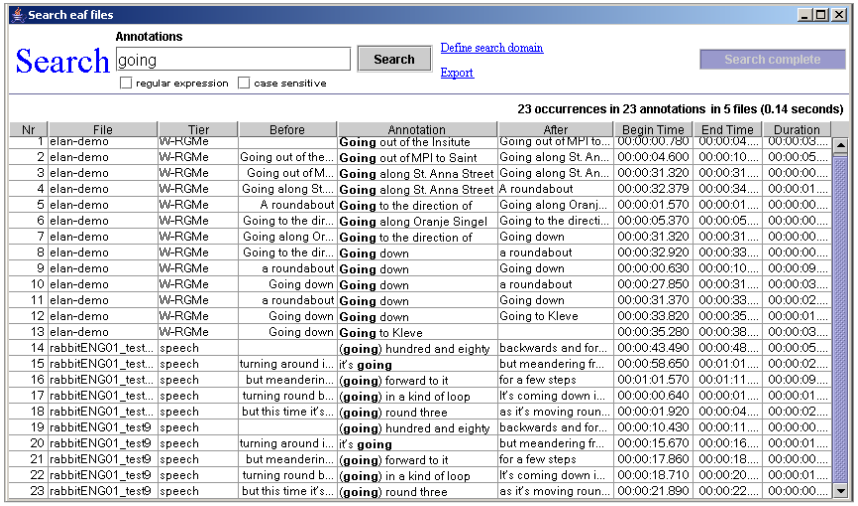
- create complex search queries
- combination of temporal and structural constraints
- use of regular expressions
- save and load queries
- export result to tab-delimited text
- search and replace



Multiple file search

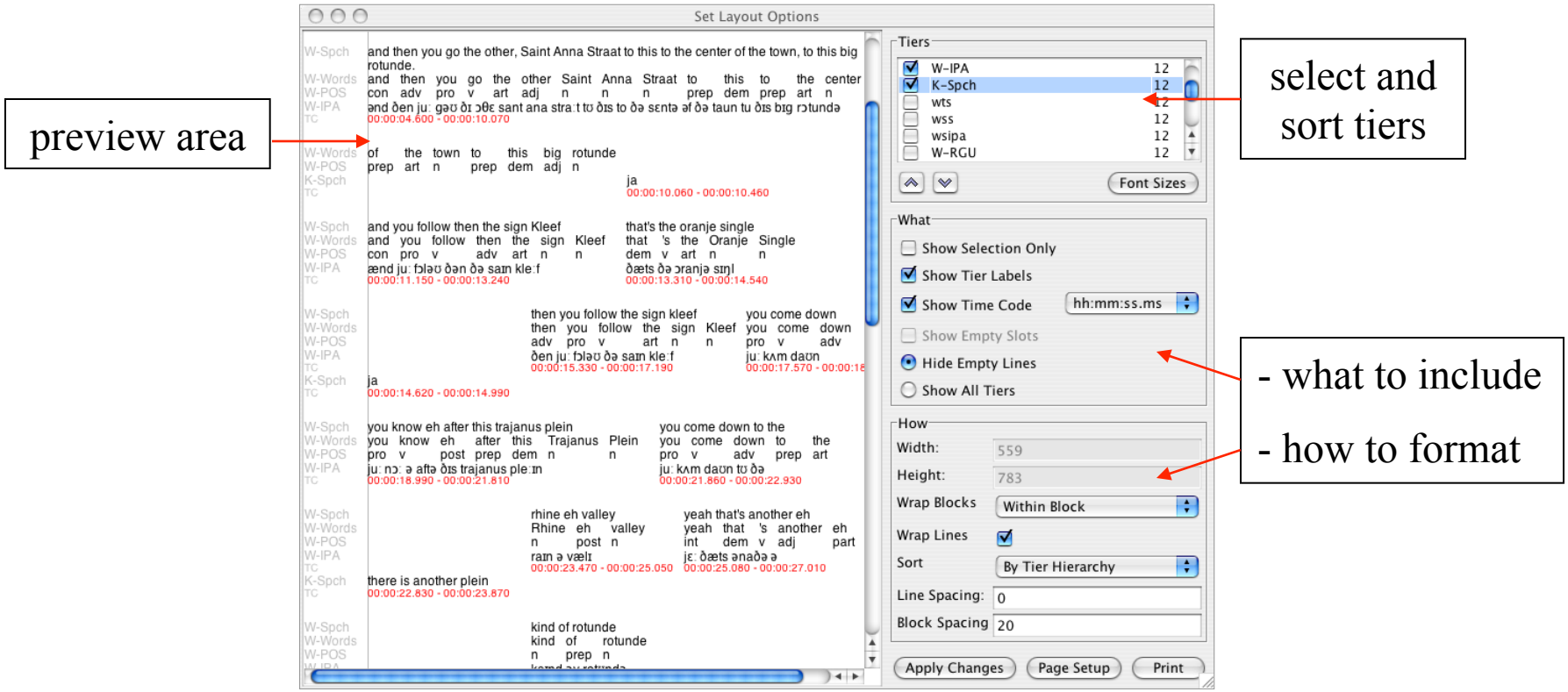
- add ELAN files to a ‘basket’
- simple and structured search: several matching types, in selected set of tiers
- concordance and frequency view
- export results to tab-delimited text file
- open a file by a click on a result





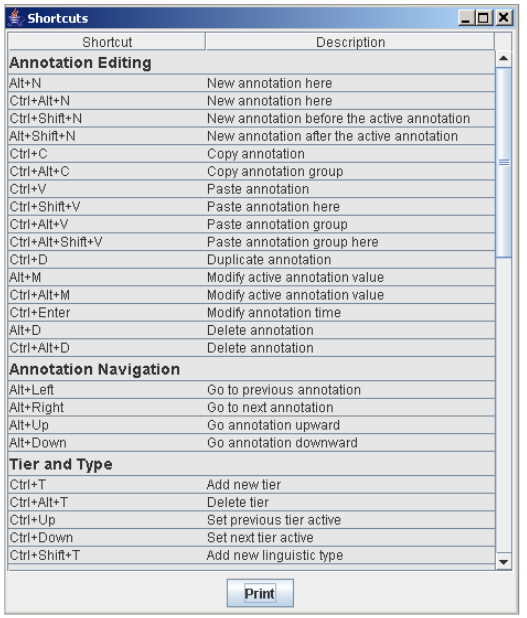
Interlinear output

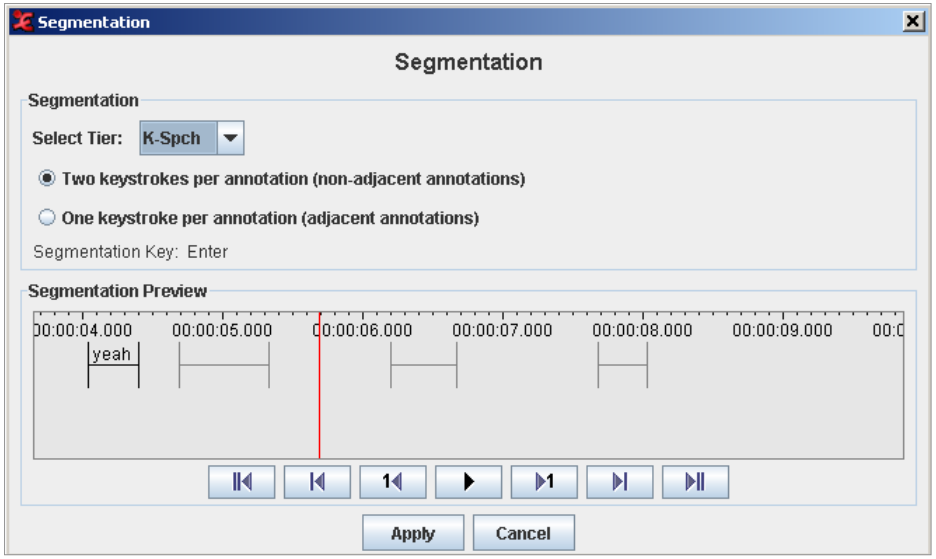
- customizable printing and export to text/html file
- selection and sorting of tiers
- selection of time code format
- line and ‘block’ wrapping



Productivity

- multiple undo / redo
- copy, duplicate, paste annotations
- keyboard shortcuts
- special annotation creation modes
  - tokenizing
  - filtering
  - semi automatic segmentation
  - annotations from overlaps

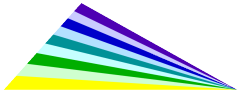




Compatibility

- Import:
  - Shoebox/Toolbox (.txt)
  - Transcriber (.trs)
  - Chat (.cha)
  - Praat (.TextGrid)

- Export:
  - Shoebox/Toolbox
  - Chat
  - annotations in tabular text format
  - Interlinear text
  - Praat
  - HTML
  - Word list, list of annotations



ELAN Linguistic Annotator

http://www.lat-mpi.eu/tools/elan