Transcribing with ELAN

ELAN (EUDICO Linguistic Annotator - http://www.lat-mpi.eu/tools/elan/) is a tier-based media annotation tool developed by the Max Planck Institute for Psycholinguistics, which can be used both for orthographic transcription, and also extensive annotation on different tiers. It can be used to annotate multiple video files, and/or an audio file.

LaBB-CAT supports using ELAN files for uploading, as long as the ELAN file contains one tier per speaker that includes orthographic transcription of their speech, like this:

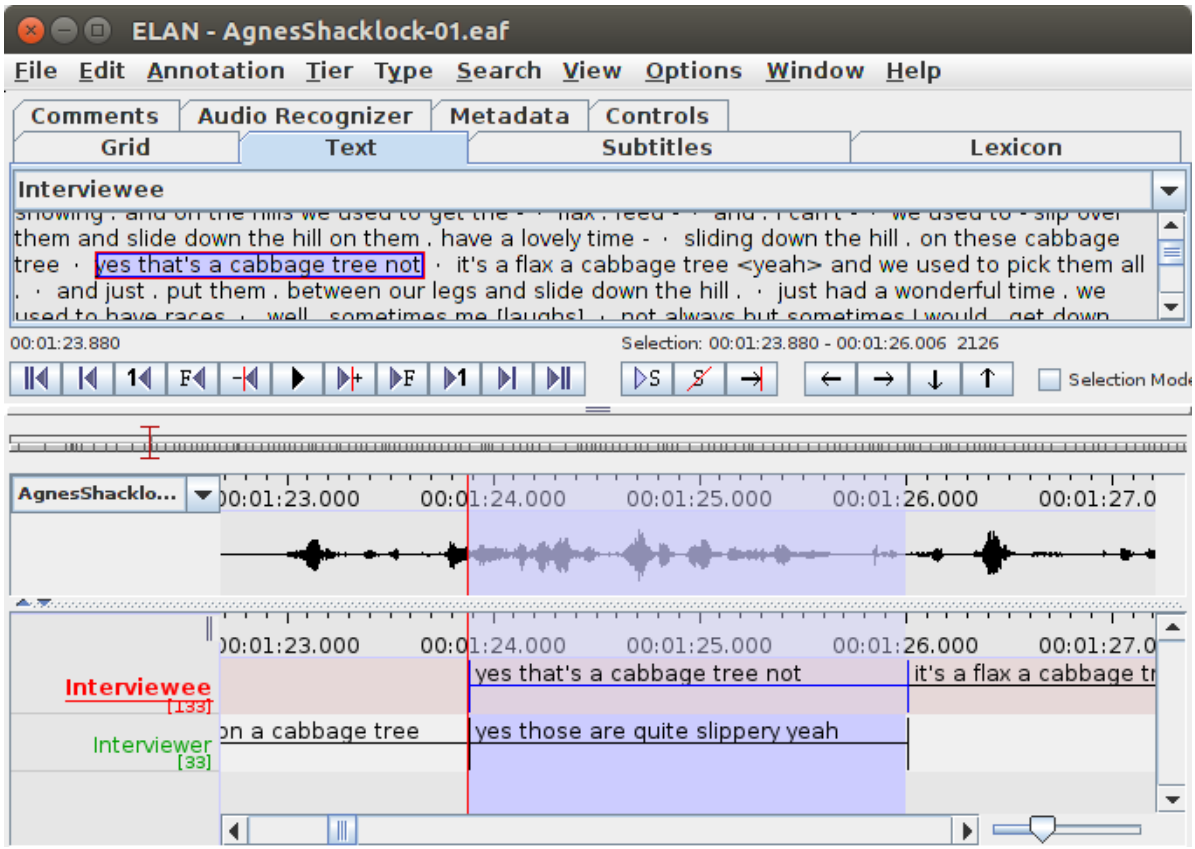

Figure 1: ELAN file with orthographic transcription, one tier per speaker

LaBB-CAT needs a mechanism for uniquely identifying the speaker of each utterance. In ELAN, the best way to achieve that is to ensure that the *Participant* attribute of each tier is set with the name (or ID) of the speaker:

If the transcript is part of multilingual corpus, you should also set the *Content Language* attribute of each tier. For more information, see Specifying Language in ELAN transcripts.

Figure 2: The Participant tier attribute is set as the speaker ID

Uploading ELAN files

To upload ELAN files, click the *upload* menu option, and then the *upload transcripts* option.

Press *Choose File* on the left and select the transcript *.eaf* file. On the right the text "ELAN EAF Transcript" should appear. Press the *Choose File* to the right of this to select the media file(s) that correspond to the transcript. Press *Upload*.

Next you will be asked to specify which ELAN tiers (listed on the left) correspond to which LaBB-CAT layers (listed on the right). Tiers that contain orthographic transcript text (as illustrated above) should be mapped to LaBB-CAT's "utterance" layer. You may also have other annotations in the transcript (e.g. a tier for noise annotations). These can be mapped to LaBB-CAT layers if a corresponding layer has already been set up in LaBB-CAT. Generally speaking, for these you'll want to create a "span" layer for time intervals, and it's best to create a LaBB-CAT layer with the "name" set to the same name as the tier, so that it's selected by default. LaBB-CAT already has a "noise" layer and a "comment" layer, so these don't need to be set up in advance.

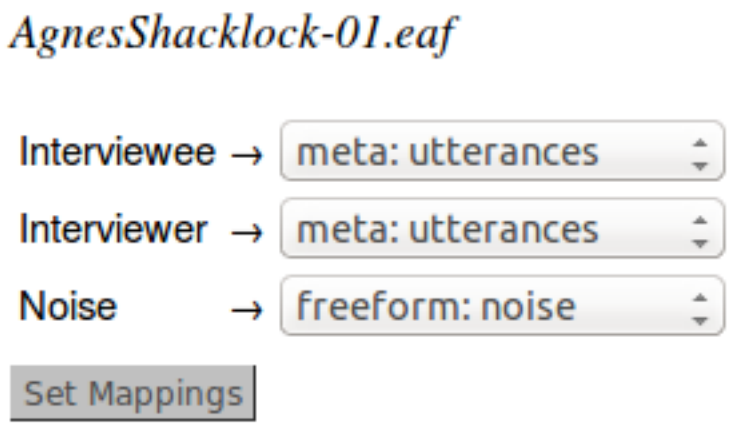

Figure 3: ELAN tiers listed on the left are mapped to LaBB-CAT layers on the right

In the illustration, the "Interverwee" and "Interviewer" ELAN tiers are mapped to the "utterances" LaBB-CAT phrase layer, and the "Noise" ELAN tier is mapped to the "noise" LaBB-CAT span layer.

When you click Set Mappings, the transcript will be imported into LaBB-CAT using the specified tier/layer correspondences.

Conventions for non-speech annotations within the transcript

ELAN has no direct mechanism for marking non-speech annotations in their position within the transcript text. However, LaBB-CAT supports the use of textual conventions in various ways to make certain annotations:

- To tag a word with its pronunciation, enter the pronunciation in square brackets, directly following the word (i.e. with no intervening space), e.g.:
  …`this was at Wingatui[wIN@tui]`…
- To tag a word with its full orthography (if the transcript doesn't include it), enter the orthography in round parentheses, directly following the word (i.e. with no intervening space), e.g.:
  …`I can't remem~(remember)`…
- To tag a word as being in a different language, enter the code CS: (for 'code switch') followed by the the ISO 639 3-letter code for the language, in square brackets, directly following the word (i.e. with no intervening space), e.g.:
  …`me mudé de New[CS:eng] Zealand[CS:eng] en 2004`…
- For longer phrases, the code-switch tag can be placed immediately before the first word and immediately after the last word, to mark those and all intervening words as being in a different language. e.g.:
  …`has a certain [CS:fre]je ne sais quoi[CS:fre] I think`…
- To insert a noise annotation within the text, enclose it in square brackets (surrounded by spaces so it's not taken as a pronunciation annotation), e.g.:
  …`sometimes me [laughs] not always but sometimes`…
- To insert a comment annotation within the text, enclose it in curly braces (surrounded by spaces), e.g.:
  …`beautifully warm {softly} but its`…

During upload, these annotations will be extracted from the transcript text and inserted into corresponding LaBB-CAT layers.