Linking to Corpus NGT movies using ANNEX

SignLinC
Deliverable 5

Funded by CLARIN-NL, 2010

Onno Crasborn
Centre for Language Studies
Radboud University Nijmegen

Version 2, 15 December 2010
Introduction

This document aims to explain how to link to segments of the Corpus NGT – and more generally to any annotated media file in the language archive of the Max Planck Institute for Psycholinguistics. These links are not trivial, and require some insight in the procedure of generating and publishing the annotations. The reason for this is that the links do not directly refer to the movies, but rather to the annotation documents that in turn are linked to the movies. What will be presented to the user is a combination of an annotation document and the movie(s) and/or audio files that are linked to it. This annotation file need not actually be filled. In fact, at the time of writing, the majority of sessions of the Corpus NGT have not yet been annotated at all.

This document starts with a characterisation of the whole workflow and the different tools that are involved. Next, the archive of the Corpus NGT and the published annotations are described. Subsequently, the creation of links using the ANNEX tool is explained.

In addition to linking to annotations that are opened in a web browser using the ANNEX web tool, it is also possible to embed a ‘mini player’ in a web page that displays a video with one or more annotated tiers. In the former way, the whole annotation file is displayed and available for inspection to the user. By using an embedded mini player, authors can restrict the display of an annotation file to just a specific time span. This embedding is described in the last section of the document.

Both of the features that are described in this document were not yet included in the latest version of the manual (July 2009) at the time of writing of this document. The manuals will be updated in due course.

Overview of workflow: from movie to annotation and back again

Movie clips (preferably using open formats like MPEG-1 or MPEG-2) can be annotated with the stand-alone tool ELAN. The resulting annotations are stored in a "eaf" file (ELAN Annotation Format), while the display preferences for the file in ELAN are stored in a preferences file (extension "pfsx").

Picture of workflow:

Movie (MPEG-1, MPEG-2)
ELAN
annotation file, preferences file
LAMUS, AMS
server; streaming movie
ANNEX
create link to annotations
miniplayer
Web browser
The Corpus NGT in the MPI archive

The MPI language archive contains the whole Corpus NGT, including both media files (movies and audio files with the voice over of interpreters) and annotation files. The media files are included in two formats: the MPEG-1 versions (extension .mpg) are used for annotation with the stand-alone tool ELAN, while the MPEG-4 versions (extension .mp4) are files that can be streamed for viewing in a browser window with ANNEX, the online sibling of ELAN.

The corpus is located in the Sign Language branch of the corpus. It is structured in three parallel directories, that all include the same sessions. One can thus browse by recording session (Recordings), by type of content (Data Type), or by Region (which in turn is sorted into Data Type at a lower level). The following screen shots illustrate this organisation.

Upon expanding a subnode, one gets to see one or more 'sessions', which are the bundles of media files and annotation files. The node with the green bag itself (the metadata session) contains the administrative information pertaining to the whole session. Upon selecting it, the metadata can be inspected in the right pane of the browser. Expanding the session leads to a list of resources:
Upon selecting a resource, one gets to see some of the properties of the file on the right-hand side. Most importantly in the present context is the accessibility information: if the file is accessible for user ‘anonymous’, this means that everyone can see the file, without having to log in. This makes the file suited for linking to in any type of document. EAF files and movie files in the Corpus NGT always have the same accessibility status, so it suffices to inspect the status of the EAF file in this corpus.

The annotation documents (extension .eaf) at the moment only include glosses for the left and right hand of each signer (so four different tiers), and only for a small subset of all 2375 movie fragments. All fragments have an EAF file linked to it, however, so that they can be inspected with ANNEX. How to do this is the topic of the following section.

In April 2011, an update of the Corpus NGT annotations will be published on the corpus server. This update will include the following features:

- A substantial expansion of the clips that have been glossed
- Major revisions in the glossing conventions, adopting ID-glosses
- Narrow translations for a subset of the corpus, and a sentence-level segmentation (without translation) for an additional subset
From the corpus to ANNEX

ANNEX can only be used with EAF files that are in the MPI archive: it cannot open local annotation documents. Thus, the simplest route to open ANNEX is to go to a specific annotation document in the archive, and select "View node" from the contextual menu:

A new web browser window will open displaying the available tiers and annotations (if any). Because there are multiple MPEG-4 movies linked to each annotation document, the user is first asked to make a selection:

After confirming the selection, the selected video will appear in a window that resembles the ELAN interface:
The recordings for the Corpus NGT featured one camera on each signer in the dialogue. These are available as separate clips (having the participant code (S005 for example) and the camera view (b for upper body) in the file name. However, there is also a combined clip that only includes the session number (CNGT0055.mp4, in the above screen shot). For most uses, this view is to be preferred. Should the resulting video window be too small, one can reopen the annotation document and choose for the video with only one signer. The interface and related terminology for ANNEX are quite similar to ELAN. For detailed information, please see the ANNEX manual online, or the PDF version.

Linking to the Corpus NGT using ANNEX

At the moment, there is not yet a standard function in ANNEX that generates a URL that would lead a user to a specific time segment of a specific EAF document. This function is in the making. In the meantime, some simple hand work is needed to generate this URL:

1. Copy the URL of the ANNEX document that appears in the browser. (Note that this URL can be very long, and not something a user would ever want to type in.)
2. Add &time=X, where X is the time in milliseconds of the start of the segment
3. Add &duration=Y, where Y is the duration of the selection that one wants to include.

This results in a URL that may look like this: http://corpus1.mpi.nl/ds/annex/protected/no_auth.jsp?nodeid=MPI654121%23&time=20000&duration=3500.

Clicking this link will lead to a new browser window that looks like this:
The user can now click the "Play Selection" button to play back just the fragment that you have pre-selected. At the same time, the user is free to look at the whole video clip and all associated annotations in the EAF document.

Some further options in addition to the start time and duration can be specified in the URL as well:

4. Select one of the six available view types to be shown as initial view type: text, grid, subtitle, timeline, waveform, combined view.
   a. Example: &Type=timeline

5. Select one or more view types to be hidden right from the start.
   a. Hide one view type (example): hideview=timeline
   b. Hide several view types, separating them by @. For example: hideview=text@combined@grid

Should one want to present a more restricted view of the annotations and media files, there is the alternative of embedding a 'mini ANNEX' in an existing web page. The next section explains how to do this.

**Embedding a mini player with a segment of the Corpus NGT using ANNEX**

Web masters can embed an annotation document as a flash object in a web page. From ANNEX, clicking the “Embed” button at the top leads to a window where one can specify some appearance features. The <DIV> tag that is generated can then be copied and pasted into an HTML document.

The editor has to select beforehand how the annotations are presented: as running text or as subtitles. After having chosen one of the available streaming media files (in case there are several), the user can then select two tiers to be displayed in the pre-defined view below the video.
In the case of the Corpus NGT movies that contain both signers in one MPEG-4 movie, the resulting view is rather small. In the current version, it is not possible to detach the media window and resize it. An alternative is to click “Start full ANNEX” at the top right of the window. This opens a new browser window with the regular ANNEX features, including the option to detach the video window, making it a bit larger.

Conclusion

Although there are currently two different ways of linking to the Corpus NGT directly from the archive, it will be clear that the available features and flexibility are limited. ANNEX will continue to be developed in the coming years, and feedback on the functioning and features of ANNEX is much appreciated by the development team. Using the ANNEX forum online will help the development and also support other users.

For using the Corpus NGT, there is also the option of downloading the MPEG-1 or MPEG-4 files from the archive, and including (segments of) those in web pages or other products. In this case, special care has to be taken that the license conditions of the corpus (Creative Commons BY-NC-SA 3.0) are respected. For research and other advanced use, downloading the MPEG-1 files together with the EAF files allow users to use ELAN to not only view but also add annotations.

Useful references

Studies, Radboud University Nijmegen. URL: [http://www.let.ru.nl/corpusngt](http://www.let.ru.nl/corpusngt)

**Links**

- **MPI Archive**: [http://corpus1.mpi.nl](http://corpus1.mpi.nl)
- **Corpus NGT (scientific)**: [http://www.ru.nl/corpusngtuk](http://www.ru.nl/corpusngtuk)
- **Corpus NGT (public, NL)**: [http://www.ru.nl/corpusngt](http://www.ru.nl/corpusngt)
- **Corpus NGT (public, DE)**: [http://www.ru.nl/corpusngtde](http://www.ru.nl/corpusngtde)