

## Using ELAN for annotating sign language corpora in a team setting

Onno Crasborn<sup>a</sup> & Han Sloetjes<sup>b</sup>

<sup>a</sup>Radboud University Nijmegen, Centre for Language Studies, PO Box 9103, NL-6500 HD Nijmegen, The Netherlands. E-mail: o.crasborn@let.ru.nl

<sup>b</sup>Max Planck Institute for Psycholinguistics, Wundtlaan 1, Nijmegen, the Netherlands. E-mail: han.sloetjes@mpi.nl

### Abstract

ELAN is a multimedia annotation tool that is employed in many sign language corpus projects. It is a standalone desktop application that, like many other desktop applications, principally is a single user, document oriented application. In many scenarios this is still perfectly satisfactory but in large-scale corpus projects, involving many collaborators who are working on the same documents, the problem arises of how to resolve edit conflicts and how to prevent undesirable modifications to parts of the document. The Corpus NGT project is such a project and this paper describes the challenges that arose in the process of its creation as well as in the exploitation of this large collection of annotation documents. It outlines recent and possible future development of ELAN and alternate solutions that have been explored and applied.

### 1. The problem

ELAN is a free, multimodal annotation tool for digital audio and video media. It supports multileveled transcription of up to six synchronized video files per annotation document. The documents are stored as XML (Extensible Markup Language)<sup>1</sup>, in its own EAF file format. Over the years, the facilities for working with multiple files have gradually increased. However, in most respects ELAN still assumes that there is a single user for those files, or that users work on the data one at a time. This situation raises several challenges in the creation of large collections of annotation documents that are jointly used by researchers working in a team, as in the case of the development and use of signed language corpora.

This paper characterises several of those challenges as they arose in the creation of the *Corpus NGT* and its subsequent exploitation for research. It shows how on the one hand this has steered the recent development of ELAN, and on the other hand complementary solutions have been found that address the complex situation of teamwork on a large set of files. It concludes by suggesting several areas for possible future development of ELAN.

### 2. Background

#### 2.1 ELAN<sup>2</sup>

ELAN has a development history of more than 10 years. The software followed the Mac-only application MediaTagger and was called EUDICO in its earliest versions, and it arose from a European project of the latter name.<sup>3</sup> The initial set of client-server based viewer applications that were developed in that project, gradually merged into a single standalone annotation editor.

ELAN has originally been, and in fact still is, strongly oriented towards a setting where single users are working on a relatively small number of annotation documents. Like many other desktop applications, and this is probably

true for a majority of them, ELAN assumes that there is only one user at a time working on a document.

At the start of the 21<sup>st</sup> century, some users expressed their wish to be able to work on annotation documents collaboratively. This led to the implementation of the onsets of a Peer-to-Peer (P2P) based solution for simultaneous, collaborative annotation (Brugman, Crasborn & Russel 2004). In this approach, team members and/or other collaborators are working together at the same time on the same document. Crucial is that the collaborators don't have to be at the same site, sitting at the same workstation. This solution has been implemented and tested up to the demonstration phase, but has never been finalised.

A disadvantage, or at least a limitation, of the above P2P type of collaboration is that the annotators need to be available at the same moment and need to be focussing on the same phenomenon. But in many team situations this is not the most suitable form of collaboration, e.g. in projects where most annotators have specialised into studying a particular kind of phenomena and are working on different tiers in different sections of the media file. One way to handle this, at least in theory, is to let each annotator work in a separate file referring to the same media file(s) and merge all these transcriptions in the end into one complete transcription file using ELAN's "Merge Transcriptions" function. In practice however, this workflow often is not realistic, if only because there is no apparent "end" to the annotation work; it is often not possible to decide when a certain part of the work is finished, and making modifications to a part of the annotation might necessitate re-merging of files. And in some cases it is useful to have the information from annotations on other tiers at hand during the annotation phase (although the opposite can be true as well).

In sections 3 and 4 we describe a combination of solutions that have been created, which consist of a combination of enhancements to ELAN and local solutions for the work with the specific collection that will first be described in the following section.

<sup>1</sup> <http://www.w3.org/standards/xml/>

<sup>2</sup> <http://www.lat-mpi.eu/tools/elan/>

<sup>3</sup> <http://www.mpi.nl/world/tg/lapp/eudico/eudico.html>

## 2.2 The Corpus NGT<sup>4</sup>

The Corpus NGT is a collection of almost 72 hours of dialogues of 92 different signers for whom NGT is the first language (Crasborn, Zwitterlood & Ros, 2008; Crasborn & Zwitterlood 2008). The recordings for the corpus were created between 2006 and 2008, and the first release of the videos with some initial gloss annotations was published as open content in December 2008. Over 15% of this material received a voice-over from sign language interpreters. A second release of the annotation files including a much larger set of ID-glosses (Johnston 2008) and some sentence-level translations will be published in 2011.

Aside from this publication for linguists as part of the MPI corpus archive, a public version of the data have been made as streaming media in early 2010. The public web site includes a presentation of the data for Deaf people, second language learners, and any interested party. The web site has been translated to German, and an English and NGT version are being planned.

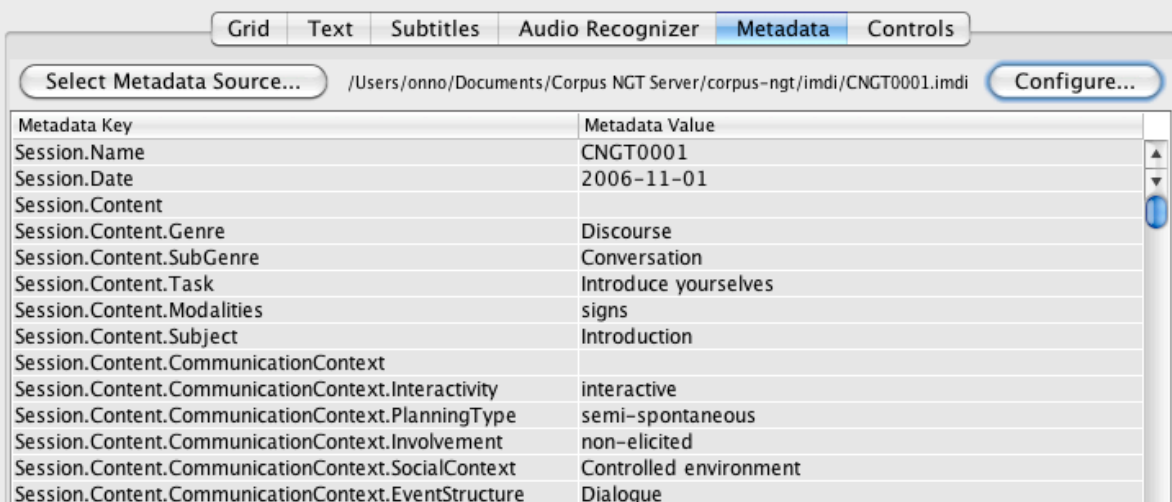
Since its original publication, the 2375 sessions in the Corpus NGT have been used for various research projects. For a project on sign language recognition (SignSpeak), additional gloss annotations are being added and the glosses are being revised to adhere to a more strict ‘one manual form, one gloss’ rule, termed ID-glosses by Johnston (2008). Moreover, for a variety of research projects at Radboud University, many new annotation levels (tiers) have been added. A total of seven researchers and four research assistants regularly add annotations to the corpus now, and perform increasingly complex searches.

### 3. Working with large sets of annotation documents

The creation of the Corpus NGT involved the segmentation of the data into 2375 parts, each consisting of one annotation file and a number of media files linked to it. Even with a much smaller number of files, it would not be realistic to want to process documents one-by-one: searching or adding tiers only in open documents would not be realistic and would lead to unsystematic files and annotations. For this reason, the Corpus NGT project contributed to the design and implementation of several new functions in ELAN.

The key development in this area was the creation of a link between the metadata descriptions of corpora and the annotation documents. Although ELAN stores some metadata properties of individual annotation documents (such as the ‘Author’ of a document and the ‘Annotator’ of a tier), metadata typically transcend the level of an individual annotation document, classifying sets of documents as sharing the same signers or the same content type or register. Until now, the metadata information that is stored in IMDI files was not accessible from within ELAN. For a search across multiple files with metadata property X, one would have to manually create a domain by selecting annotation documents corresponding to that metadata property one by one in a file selection dialogue, where this information would have to come from another source (such as the IMDI files or another database with the metadata information). Similarly, in order to quickly inspect from which region a participant in the media file comes, one would have to look up the session number in the metadata records.

The first addition that was created to facilitate access to metadata was the creation of a new tab pane in the top right hand part of the ELAN interface. Next to the Grid, Text, Subtitle and Controls pane, a Metadata pane has been created in which the user can select an IMDI file and the fields to be displayed in a table view (Figure 1) or in a



The screenshot shows the ELAN interface with the 'Metadata' tab selected. A 'Select Metadata Source...' button is active, showing the path: /Users/onno/Documents/Corpus NGT Server/corpus-ngt/imdi/CNGT0001.imdi. A 'Configure...' button is also visible. Below is a table with two columns: 'Metadata Key' and 'Metadata Value'.

Metadata Key	Metadata Value
Session.Name	CNGT0001
Session.Date	2006-11-01
Session.Content	
Session.Content.Genre	Discourse
Session.Content.SubGenre	Conversation
Session.Content.Task	Introduce yourselves
Session.Content.Modalities	signs
Session.Content.Subject	Introduction
Session.Content.CommunicationContext	
Session.Content.CommunicationContext.Interactivity	interactive
Session.Content.CommunicationContext.PlanningType	semi-spontaneous
Session.Content.CommunicationContext.Involvement	non-elicited
Session.Content.CommunicationContext.SocialContext	Controlled environment
Session.Content.CommunicationContext.EventStructure	Dialogue

Figure 1. The Metadata pane in ELAN displays a selection of metadata properties from an IMDI file.

<sup>4</sup> <http://www.ru.nl/corpusngtuk>

tree view.

Secondly, multiple file searches were enhanced so that they can make use of the output of a search in the IMDI Browser. Thus, in a two-step process one can first search for metadata characteristics and then use the outcome of that search for annotation searches in ELAN. To this end, the IMDI Browser was adapted so that it would be possible to save search results in a file that can then be read by ELAN. The selection of the IMDI file and the specific metadata fields to display is stored in the preferences file.

Most of the available multiple file processes, such as

that allows adding, changing and deleting tiers and linguistic types in multiple documents. Here too, the user can make a selection of a domain to modify and store that domain for later use. The implementation offers a tabular overview of the different tiers and tier properties (Linguistic Type, Annotator, Participant) that are used by all the files in the set, which can help to keep a corpus organised. As users are free to add new tiers to and modify existing ones in any document in a corpus collection, they can also create inconsistencies. These can be easily spotted in the *Multiple File Editor* interface (Figure 2).

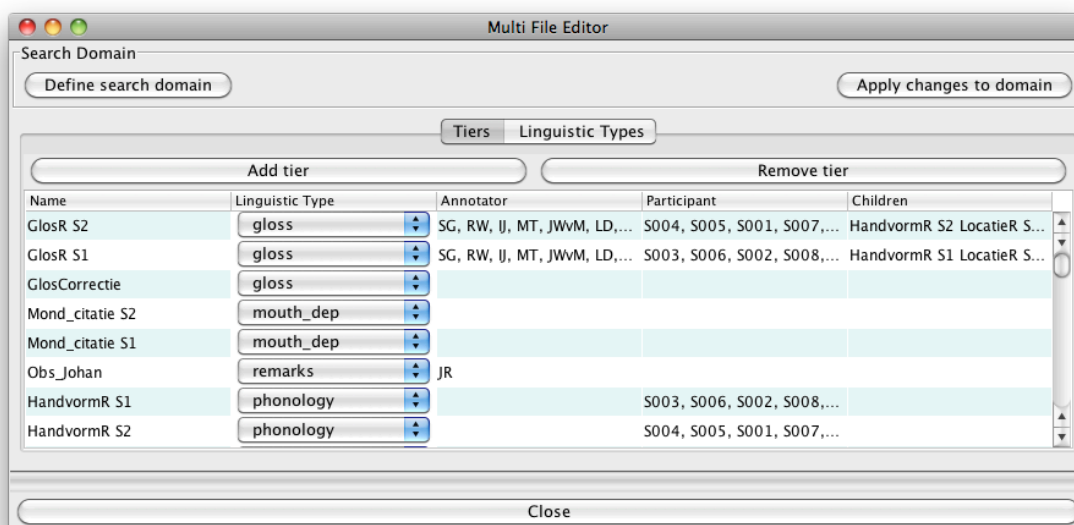


Figure 2. The Multiple Files edit function gives an insightful overview of properties of tiers and their properties.

searching in multiple annotation files, are accompanied by a “domain” selection facility. Domains in this context are selections of files and folders that can be saved in ELAN and reused later. Domains can either be composed manually, by selecting files and folders in a custom file browser window, or they can be derived from an IMDI metadata search result as described above.

Actions that can be performed on multiple files now consist of the following:

- structured search
- find and replace
- generation of statistics
- new document creation based on a template and sets of media files
- annotation “scrubbing” (removal of superfluous spaces, tabs and new lines)
- export as word list, export a selection of tiers and export to tab-delimited text

For all these purposes, then, a selection can be made in the IMDI Browser so that the action would only apply to annotation documents that relate to, for instance, signers from a specific region or from a specific age group.

A special case of processing multiple files is the module

#### 4. Working with a research team on a corpus of annotation files

One of the changes in ELAN made to improve the work in a team setting has been the introduction of the ‘Annotator’ attribute in the specification of tiers. This has been added in ELAN version 3.0; at the same time a corresponding change was made in the EAF schema, in version 2.4. This attribute can be used to sort or group tiers and for creating statistics per annotator. It is expected that the existing “Compare Annotators” function will be extended to make use of this attribute. This function currently produces a rough calculation of the level of agreement between two annotators or raters.

Other tasks were not yet implemented in ELAN at the time of the construction of the Corpus NGT. One function (currently under development) was to create new EAF files based on a template and a list of media files. To facilitate the generation of new documents for the Corpus NGT, Perl scripts were written to create EAF files for a set of media files, and to create PFSX files for a folder of EAF files, based on a dummy PFSX file that was configured to meet specific needs.

In the Corpus NGT annotation documents, specific tiers have been created for exchanging information. There is a *Observations* tier per team member, in which notes for colleagues can be stored. The tier *GlossCorrection* is used for marking possible errors in the glosses, to be double-checked by a team member with that responsibility.

As ELAN is not set up as a client-server system, a solution was sought in which the annotation documents would still be stored in a central space and accessible for all team members. A satisfactory solution until now has been to use the Subversion (SVN) file versioning system, which is typically used in the context of software development in teams. There is a SVN server on the network that creates a backup of every version of every annotation document ever created. When storing a new revision of a file, annotators can add comments as to what was changed in this version of the file. Aside from the backup facility, an advantage of this system is that all users can immediately profit from new annotations as soon as they are uploaded to the server.

The downside of the versioning system is that it imposes heavy demands on the users to stick to strict workflows. Repairing conflicting versions may take quite some time. Moreover, it is not a principled solution: Subversion is really targeted at situations where the text files *themselves* are edited by users, as in software development. In the case of EAF documents, which are an instance of XML, ELAN assumes that there are no other editors of the XML code than ELAN itself, and this can make comparing conflicting versions rather hard. This is particularly so when it comes to the coding of time positions and annotation IDs.

In addition to the EAF files, the SVN server also hosts all the IMDI metadata files and one folder of PFSX files per researcher or research goal. The location of the folder with preferences files can be set in the ELAN preferences since version 3.7.2. Users can thus have access to a uniform ELAN interface for all the documents they open, irrespective of who most recently edited the document.

The applicability of preferences files has been improved by saving preferences when a template file is created. Every new annotation file based on such a template with an associated preferences file, starts with the inherited preferences settings.

## 5. Areas of further development

The development of a more systematic use of the concept ‘user’ could further facilitate the use of ELAN in teams. Perhaps the possibility of choosing a server-client setup where information about user actions can be systematically stored and conflicts between actions of different users can be prevented would merit consideration again. The iLex tool uses this type of design successfully.<sup>5</sup> This might entail a shift from an XML document oriented approach to a managed database oriented approach.

<sup>5</sup> <http://www.sign-lang.uni-hamburg.de/ilex>

There are a number of issues and wishes, brought forward by several user groups, which are seemingly related to the issues discussed in this paper:

- In team settings a need has emerged to “write protect” certain parts of the document for all or most of the annotators.
- Documents that were created based on a template file, can easily become inconsistent when tiers are renamed or deleted.
- Support for a “stand off” treatment of tiers in different transcription files. The tiers of only one of the files should be editable; the other tiers should be read only.

Finding a way to converge these issues and develop, if possible, a single solution is one of the challenges for future developments.

## 6. Acknowledgements

The software development and the writing of this paper was made possible by grants from the Netherlands Organisation for Scientific Research (NWO, grants 380-70-008 and 276-70-012) and the European Research Council (ERC Starting Researcher Grant 210373 awarded to Onno Crasborn). Developers who contributed to the improvements in ELAN that are described in this paper: Mark Blokpoel (RU), Albert Russel (MPI), Eric Auer (MPI), Alexander Koenig (MPI).

## 7. References

- Brugman, H., Crasborn, O., Russel, A. (2004) Collaborative Annotation of Sign Language Data with Peer-to-Peer Technology. In: *Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation*. M.T. Lino et al., eds. Pp. 213-216.
- Crasborn, O. & I. Zwitterlood (2008) The Corpus NGT: an online corpus for professionals and laymen, In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd, eds. ELDA, Paris, pp 44-49.
- Crasborn, O., I. Zwitterlood & J. Ros (2008) The Corpus NGT. A digital open access corpus of movies and annotations of Sign Language of the Netherlands. Nijmegen: Centre for Language Studies, Radboud University Nijmegen.  
URL: <http://www.ru.nl/corpusngtuk/>
- Johnston, T. (2008) Corpus linguistics and signed languages: no lemmata, no corpus. In: *Construction and Exploitation of Sign Language Corpora. 3rd Workshop on the Representation and Processing of Sign Languages*, O. Crasborn, T. Hanke, E. Efthimiou, I. Zwitterlood & E. Thoutenhoofd, eds. ELDA, Paris, pp. 82-87.
- Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers*, 41(3), 841-849.