Enhanced ELAN functionality for sign language corpora

Onno Crasborn, Han Sloetjes

Department of Linguistics, Radboud University Nijmegen
PO Box 9103, NL-6500 HD Nijmegen, The Netherlands
Max Planck Institute for Psycholinguistics
PO Box 310, 6500 AH Nijmegen, The Netherlands
E-mail: o.crasborn@let.ru.nl, han.sloetjes@mpi.nl

Abstract

The multimedia annotation tool ELAN was enhanced within the *Corpus NGT* project by a number of new and improved functions. Most of these functions were not specific to working with sign language video data, and can readily be used for other annotation purposes as well. Their direct utility for working with large amounts of annotation files during the development and use of the *Corpus NGT* project is what unites the various functions, which are described in this paper. In addition, we aim to characterise future developments that will be needed in order to work efficiently with larger amounts of annotation files, for which a closer integration with the use and display of metadata is foreseen.

1. The *Corpus NGT* project[1]

1.1 General characterisation

The *Corpus NGT* that was published in May 2008 is one of the first large corpora of (semi)spontaneous sign language use in the world, and the first to become publicly available online. It is targeted primarily at linguistic researchers, but due to its open access policy can also be used for other purposes, whether scientific, educational, or private. The corpus consists a large collection of sign language video recordings with linguistic annotations and audio translations in Dutch. Recordings were made of nearly 100 signers communicating in pairs. This resulted in 2,000 segments totaling 75 hours. The use of multiple cameras for four different angles resulted in a collection of ± 15,000 media files.

The four different angles can be displayed in sync by the ELAN annotation tool; for this purpose, an annotation file was created for every time segment. These documents were created from a template containing multiple (empty) tiers for glosses, translations and remarks. Over 160 files were actually annotated with gloss annotations on four different tiers, one for each hand of each of the two signers. In total, over 64,000 gloss annotations were added to these files. As two-handed lexical items receive a separate gloss for the left and for the right hand (each with their own alignment), the number of annotations cannot be blindly equated with the number of signs.

Further technical and linguistic information on the *Corpus NGT* can be found in Crasborn & Zwitserlood (this volume) and Crasborn (this volume), as well as on the corpus web site: www.let.ru.nl/corpusngt/. The corpus is currently hosted at the corpus server of the Max Planck Institute for Psycholinguistics, and part of their Browsable Corpus.[2]

1.2 Use of standards and tools

The *Corpus NGT* makes use of open standards for its publication, aiming to guarantee long-term availability:

- Media files conform to the various MPEG standards (MPEG-1, MPEG-2, MPEG-4), rather than popular commercial formats such as Adobe Flash video.
- Metadata descriptions are made conforming to the IMDI scheme (Wittenburg, Broeder & Sloman, 2000; IMDI Team, 2003).[3] While this format may not be used in ten years time, its widespread use in linguistics and the publication of the whole corpus as part of a larger set of IMDI corpora at the Max Planck Institute for Psycholinguistics ensures that the corpus will be part of larger conversion efforts to conform to future standards.
- The annotation files were all created with ELAN and thus conform to the specification for EAF files (Brugman & Russell 2004).[4]

2. Developments in the ELAN software

The *Corpus NGT* project involved annotating many hours of video and a large number of annotation documents. The first aim of the technological goal of software improvement in the *Corpus NGT* project was to ease annotation. A second aim was to facilitate the use of annotation documents, in its widest sense: browsing, searching, and data analysis.

The functions described in this section appeared in a series of releases between versions 2.6 and 3.4. Specifications were set up by the *Corpus NGT* project and the ELAN developers. For guidelines on how to use the functions, including the location in menus and keyboard shortcuts, we refer to the ELAN manual.[5]

2.1 Extension of the EAF specification and a change in the preferences format

- The property ‘annotator’ has been added in the specification of tiers, allowing groups of researchers to separate which tier has been filled by whom. It is expected that this property will become a selection criterion in the

[1] The *Corpus NGT* project was made possible by an investment grant from the Netherlands Organisation for Scientific Research (NWO), grant no. 380-70-008.

[2] http://corpus1.mpi.nl

[3] http://www.mpi.nl/IMDI/schemas/xsd/IMDI_3.0.xsd

[4] http://www.mpi.nl/tools/elan/EAFv2.5.xsd

[5] http://www.lat-mpi.eu/tools/elan/manual/

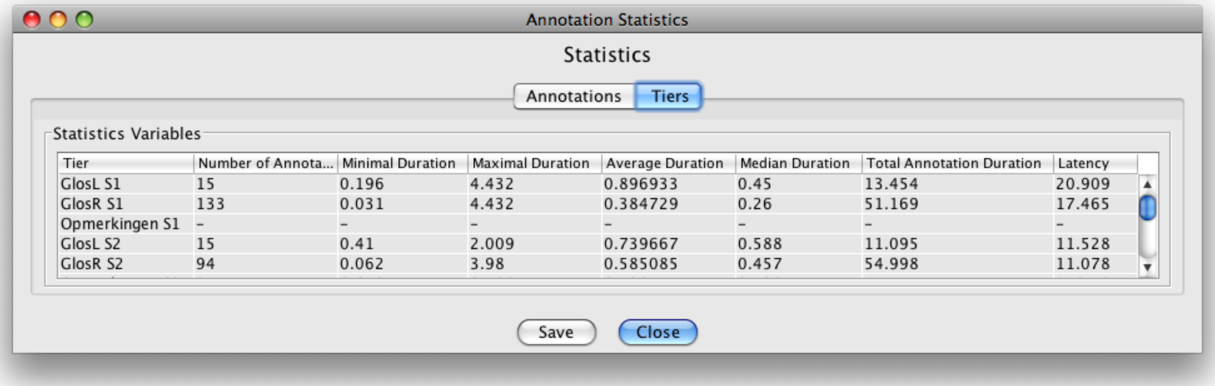
search mechanism in a future release of ELAN.

- Preferences are no longer stored in binary .pfs files, but in user-readable XML files. The current preferences settings can be exported to a file and imported in (i.e. applied to) any other document; in this way, the ordinary user without knowledge of XML can also copy settings from one document to an other. In this way, it has become easy to homogenise the layout of larger sets of ELAN documents and modify this ‘style sheet’.

2.2 New functionality

- The **‘duplicate annotation’** function was created to facilitate the glossing of two-handed signs in cases where there are separate tiers for the left and the right hand: copying an annotation to another tier saves annotators quite some time, and prevents misspellings. A disadvantage of using this function turned out to be that annotators may no longer play close attention to the timing differences between the two hands in two-handed signs. While the hands often do not start and end their articulation of a sign at the same time, the ‘duplicate annotation’ function makes it attractive to classify a sign as a phonologically two-handed form, even though the phonetic appearance can show differences between the two hands. Moreover, larger timing differences between the two hands have shown to play a role in many levels of the grammar of signed languages beyond the lexicon (Vermeerbergen, Leeson & Crasborn 2007). It will depend on the user’s research goal whether or not detailed timing differences are important to annotate correctly.

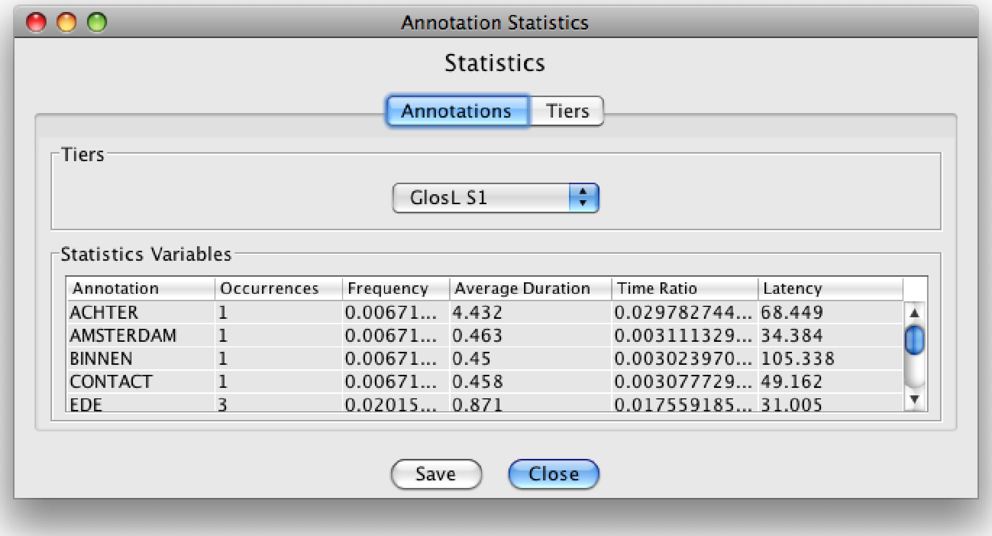
In addition to this quick annotation duplication shortcut some more generic copy and paste actions have been added. An annotation can be copied to the clipboard either as single annotation or as a group with all ‘dependent’ annotations. Pasting of an annotation or a group of annotations is not restricted to the same time segment (i.e. an annotation can be pasted at a different position in the timeline) or to the same document.



Statistics Variables

Tier	Number of Annota...	Minimal Duration	Maximal Duration	Average Duration	Median Duration	Total Annotation Duration	Latency
GlosL S1	15	0.196	4.432	0.896933	0.45	13.454	20.909
GlosR S1	133	0.031	4.432	0.384729	0.26	51.169	17.465
Opmerkingen S1	–	–	–	–	–	–	–
GlosL S2	15	0.41	2.009	0.739667	0.588	11.095	11.528
GlosR S2	94	0.062	3.98	0.585085	0.457	54.998	11.078

Figure 1. Tier statistics



Tiers

GlosL S1

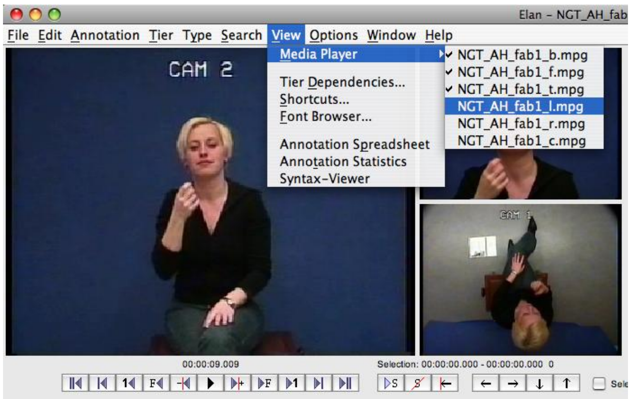
Statistics Variables

Annotation	Occurrences	Frequency	Average Duration	Time Ratio	Latency
ACHTER	1	0.00671...	4.432	0.029782744...	68.449
AMSTERDAM	1	0.00671...	0.463	0.003111329...	34.384
BINNEN	1	0.00671...	0.45	0.003023970...	105.338
CONTACT	1	0.00671...	0.458	0.003077729...	49.162
EDE	3	0.02015...	0.871	0.017559185...	31.005

Figure 2. Annotation statistics

- A new variant of **'multiple file search'** was implemented. In addition to the pre-existing 'simple text search' in multiple files, now structured searches combining search criteria on different tiers can be carried out in a subset of files that can be compiled by the user. The matching types 'exact match', 'substring match' and 'regular expression match' are available and the search can be restricted to a certain time interval. It is also possible to specify a minimal and/or maximal duration for matches.

The results can be displayed in concordance view, with a variable size of the context, or in frequency view, showing the absolute number of occurrences of each hit as well as the relative number (percentage). The results can be exported to a tab-delimited text file with multiple columns.



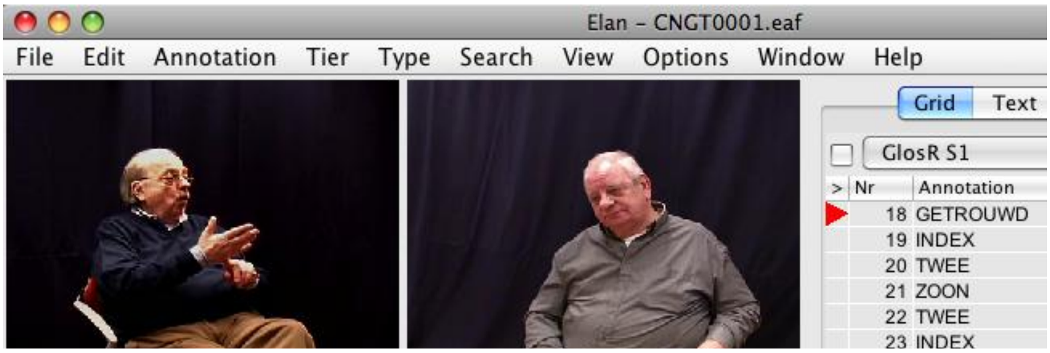
As a special case, a search for n-gram patterns can be executed, where the pattern should be found either within (multiword) annotations or over annotations on the same tier.

- The **segmentation** function was further developed so that annotations with a fixed, user definable duration can be created by a single key stroke while the media files are playing. The keystroke can either mark the beginning of an annotation or the end. Keyboard navigation through the media has been made in accordance with this function in the main window.
- A function has been added to flexibly **generate annotation content** based on a user definable prefix and an index number. Indexing can be performed on the annotations of a single tier or on those of multiple tiers.
- A panel can be displayed that lists **basic statistics** for all tiers in an annotation document (Fig. 1): the number of annotations, the minimum, maximum, average, median and total annotation duration per tier, and the latency (start time of the first annotation on that tier). This helps the user getting a better grip on the content in an annotation document and can be helpful in data analysis.

In the same window, a panel can be displayed with a list of unique annotation values for a user-selectable tier (Fig. 2): their number of occurrences and frequency as a fraction of the total number of annotations, the average duration, the time ratio, and the latency (time of first occurrence in the document).

Both panels can be saved as a text file with tab-separated lists.

- The **annotation density viewer** can now also be set to only show the distribution of annotations of a single, selectable tier. The label of a tier in the timeline viewer can optionally show the current number of annotations on that tier.
- The list of existing export options has been enriched by an option to **export a list of unique annotation values** or **a list of unique words** from multiple annotation documents. In the latter case, annotation values are tokenized into single words before evaluating their uniqueness.
- The media files that are associated to a document could already be inspected, added and removed by the 'linked files' viewer in the 'Edit' menu. Now, easy interactive **hiding and showing of any of the associated video files** is possible, without having to remove the media file association altogether (Figure 3). The maximum number of videos that can be displayed simultaneously is four. But it is possible to add more than four videos to a document and by interactively hiding or showing videos any combination of them can be shown. Temporarily hiding one or more videos can also be useful to improve


Elan – CNGT0001.eaf
File Edit Annotation Tier Type Search View Options Window Help
Grid Text
GlosR S1

>	Nr	Annotation
▶	18	GETROUWD
	19	INDEX
	20	TWEE
	21	ZOON
	22	TWEE
	23	INDEX

Figure 4. New structure of the menu bar

playback performance, especially on less powerful computers.

- A click on a video image copies the **x and y coordinates of the mouse pointer** to the clipboard. The coordinates can then be pasted into any annotation. This can be useful e.g. to record the position of body parts at various moments in time. There are three variants in the format of the coordinates. The reason for this is the ambiguity of dimension and aspect ratio in some popular media formats. As a result, media frameworks can differ in their interpretation of the video dimensions. This has to be taken into account when files are transferred between platforms, ELAN being a multi-platform application running on Windows, Mac OS X and Linux.

2.3 User interface

In addition to new functionality, a large number of user interface improvements have been implemented, including the following.

- There is an improved, more intuitive layout of the main menu bar. Due to the increase of functionality, reflected in the growth of the number of items in the menus, some menus had become overpopulated and inconvenient. The key concepts in ELAN 'Annotation', 'Tier' and 'Linguistic Type', were promoted to their own menu in the main menu bar (Figure 4).
- Many additional keyboard shortcuts have been added. The list of shortcuts is logically subdivided into groups of functionally related items and can now be printed.
- A recent files list has been added.
- Easy keyboard navigation through the group of opened documents/windows is now possible.
- There has been a subtle change in the background of the timeline viewer, facilitating the perception of the distinction between the different tiers by the use of lighter and darker horizontal bars (a 'zebra' pattern; Figure 5).
- With the use of a new preferences system in version 3, users can now set the **colour of tier labels** in the timeline viewer, thus allowing the visual grouping of related tiers in documents containing many tiers by setting the same colour for multiple tiers (as can also be seen in Figure 5). It is also possible to select a preferred font per tier; a Font Browser is included to simplify selection of a suitable font.

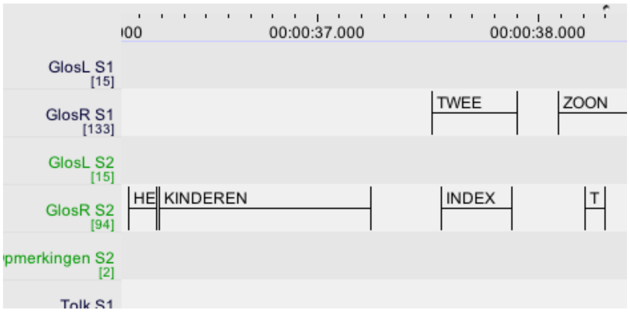


Figure 5. Striped background of the timeline viewer; tier labels with identical colours

- Previously, video windows could only be enlarged (e.g. to view details) or reduced (e.g. to have more screen space for other viewers) by detaching video windows one by one, and adapting the size of each. A function has been added whereby the video windows that are displayed can all be made smaller or larger by dragging a double arrow on the right hand side of the window above the time line viewer. All other viewers automatically resize accordingly, to keep the size of the window constant.

3. Future developments

Within ongoing projects, several new needs have become clear which all relate to the fact that suddenly the number of annotation documents that linguists can work with has increased from a small number that one can handle by hand to a huge number (around 2,000 for the *Corpus NGT*). Special attention is needed to keeping the collection well-organised (section 3.1) and to trying to use the available IMDI metadata descriptions to get a grip on the data (section 3.2). In addition, collaborative work with ELAN files is discussed in section 3.3.

3.1 Manipulating collections of files

Although enhanced search functionalities and templates facilitate working with multiple ELAN documents, it is not yet possible to 'manage' a set of ELAN files systematically in any way. For the specific files and needs of the *Corpus NGT*, Perl scripts were developed in order to add tiers and linguistic types to a set of documents, to change annotation values in multiple documents, and to generate ELAN and preferences files on the basis of a set of media files and existent annotation and preferences files.

For future users, it would be beneficial if such kind of functionality would become available in a more stable and integrated way, whether in ELAN, in the IMDI Browser, or in a stand-alone tool that can manage EAF files.

3.2 Use and display of IMDI metadata in ELAN

Current collaboration between the ELAN developers at the Max Planck Institute for Psycholinguistics and the sign language researchers at Radboud University are targeted at enhancing search facilities and facilitating team work between researchers using large language corpora containing ELAN documents.

Currently, annotation files that are included in an IMDI corpus can be searched using ANNEX, the web interface to annotation documents[6], after a subset of metadata sessions has been selected through an IMDI search. For example, one can first search for all IMDI sessions that include male signers, and then search in all EAF files that are linked to the resulting set of IMDI sessions. In this way, metadata categories and annotations can be combined.

However, currently, ANNEX cannot be used for many tasks: annotations cannot be added, edited or deleted, and the synchronous playback of multiple video streams is not

[6] http://www.lat-mpi.eu/tools/annex/

accurate. A separate two-step search is thus being developed for local corpora and the stand-alone version of the IMDI Browser.

Searching is a useful way to combine data and metadata categories, but it implies that one knows what one is looking for. Browsing through an annotation document can also be useful for many types of research, but in that case, metadata information is not available unless one knows it by heart. While the gender of the signer/speaker can be easily established by looking at the video, this does not hold for many other categories: regional or dialect background of the participant, deafness, precise age, recording date, etc. It is therefore important to have quick access to the metadata information linked to an annotation document. This requires that an IMDI metadata description is present, and that the EAF file is linked to the IMDI session. Currently, different ways of displaying metadata information in ELAN are being investigated. Some form will be available in a future version of ELAN in 2008.

3.3 Collaborative annotation

Larger collections of files are typically used not by single researchers but by research groups, and stored not on a local drive but on network drives or integrated in a corpus. This requires some type of systematic 'collaborative annotation' to ensure that changes made by one person are also available to others. Moreover, one could imagine that people add different values to annotations, that are simultaneously present and can be compared. This would be particularly useful for different translations or analyses of the same construction. Brugman et al. (2004) already discussed ways in which users at different locations look at and edit annotation documents together. We expect this concept to be further developed in the future.

4. Conclusion

A corpus building project like the present one clearly provides a fruitful collaboration between software developers and the users of the software. Although the fact that the *Corpus NGT* project was carried out on the same campus as the Max Planck Institute for Psycholinguistics facilitated collaboration, one can certainly imagine that future corpus projects reserve budget for similar collaborations between software developers and linguists. In this way, relatively small software tools can gradually be developed to match the needs of large groups of users.

5. References

Brugman, H., O. Crasborn & A. Russell (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.

Brugman, H. & A. Russell (2004). Annotating Multi-media / Multi-modal resources with ELAN. In: Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation.

IMDI Team (2003), IMDI Metadata Elements for Session Descriptions, Version 3.0.4, MPI Nijmegen. http://www.mpi.nl/IMDI/documents/Proposals/IMDI_MetaData_3.0.4.pdf

Vermeerbergen, M., L. Leeson & O. Crasborn (Eds.). (2007). Simultaneity in signed languages: form and function. Amsterdam: John Benjamins.

Wittenburg, P., D. Broeder & B. Sloman (2000), EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper. LREC 2000 Workshop, Athens. http://www.mpi.nl/IMDI/documents/Proposals/white_paper_11.pdf