Annotating Multi-media / Multi-modal resources with ELAN

Hennie Brugman, Albert Russel

Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen
{Hennie.Brugman, Albert.Russel}@mpi.nl

Abstract

This paper shows the actual state of development of the manual annotation tool ELAN. It presents usage requirements from three different groups of users and how one annotation model and a number of generic design principles guided the choices made during the development process of ELAN.

Introduction

At the Max-Planck-Institute for Psycholinguistics[1] (MPI) software development on annotation tools for the manual annotation of multimedia data has been going on since the early 90's. Over this decade there have been large changes in enabling technology and insights in the nature of linguistic annotation. Media frameworks for the handling of digital audio and especially digital video files have matured, as has media streaming technology. XML has come to existence and has become highly relevant in a short time. Rendering and input of Unicode characters is now commonplace.

Simultaneously, users made experiences with the first generation of video annotation tools and became aware of and got used to these new technologies. From this a new set of requirements arose.

Finally, annotation tool builders are better aware of each other's approaches, annotation models and annotation document formats. Clearly convergence is going on, leading to easier exchange of data between annotation tools. An important role in this process was played by the paper by (Bird & Liberman, 2001) that introduced Annotation Graphs. We are closely watching and trying to participate in standards initiatives, as for example ISO TC37/SC4.

The first video annotation tool developed at the MPI was MediaTagger, a QuickTime based application that runs only on pre-OS X Macintoshes. It started as a first attempt to exploit the QuickTime Movie data structure, and especially it's text tracks, as an informal model for linguistic annotation. Since then several new formal models where made, each one building on the experiences of the previous ones and considering new user requirements. The formal modeling languages that were used are Entity-Relationship diagrams and UML. A detailed presentation and evaluation of these models can be found in (Brugman & Wittenburg, 2001).

The next chapters will discuss the requirements of several different groups of users and describe the latest state of ELAN functionality. We will then present our model for annotation in some detail and show how we can cover the needs of very different user groups with one relatively simple model. In the discussions plans for future development will presented.

User requirements

ELAN is developed with a number of different user groups in mind. These users are situated both within the MPI and, in an increasing number of cases, outside the MPI. Often they are participating in externally funded projects (DoBeS[2], ECHO[3]). We will discuss the main requirements per group, although there is of course a substantial overlap between each group's needs.

Linguistic research

For many linguists one of the first steps in their research is the creation of an orthographic or phonetic transcription of some recorded event or experiment. In an iterative process they add more and more analytic layers of annotation to this transcription. These additional layers typically do not annotate the primary (speech) signal anymore, but refer to previously added annotations.

Layers that are added later are typically connected to already existing layers in increasingly complex referential structures. Orthography or phonetic transcription is linked directly to media time intervals in the primary signal utterance-by-utterance or phrase-by-phrase. Words are either ordered decompositions of these utterances, or are linked to media time themselves. Morphemes are ordered sequences of annotations that are symbolically linked to words. Part-of-speech annotations can refer to either words or morphemes. Structurally more complex annotation layers are recursive trees (syntax), non-contiguous annotations (co-reference), or annotations that refer to other annotations across several layers (general comments). All of these structural requirements are covered by a relatively simple and elegant annotation model called Abstract Corpus Model (ACM). ACM will be discussed in more detail in a next chapter.

An additional requirement from linguists is support for import and export of legacy annotation formats, the most important ones being the Childes format CHAT (MacWhinney) and Shoebox[4].

With respect to searching, linguists are typically interested in locating patterns on specific tiers, with the possibility to relate different patterns by means of a distance specified in milliseconds or in number of annotations on some tier. Results can be visualized in the context of their containing documents or in concordance-like representations, or they

[1] http://www.mpi.nl

[2] http://www.mpi.nl/DOBES

[3] http://www.mpi.nl/ECHO

[4] http://www.sil.org

can be the input for modules that calculate specific statistical or linguistic measures.

Documentation of Endangered Languages

One of the main application areas for ELAN is the documentation of endangered languages, both by MPI researchers and by field teams participating in the DoBeS project (Dokumentation Bedrohter Sprachen – Documentation of Endangered Languages), funded by the Volkswagen Stiftung[5].

Since one of the main components of language documentation is the result of linguistic research, all linguistic requirements hold here as well. With respect to complex annotation structures this is illustrated very clearly by the Advanced Glossing proposal (Drude & Lieb, 2002) that was made in the context of the DoBeS program. More than in the case of general linguistics, the support for entry and rendering of Unicode characters is required.

With respect to legacy formats *interlinear text*[6] in a range of document formats and with a number of proprietary and often undocumented conventions is widely used. Conversion of such texts to archival formats is required. For a good description of requirements for archive formats, see (Bird & Simons, 2003).

For a complete language documentation of some linguistic event or text it is also necessary to document used terminology (such as for example used tag sets) in an archival format.

Other important products of language documentation are lexica. The existence of lexica imposes additional requirements on ELAN. First, users want to add lexicon entries from the context of an annotation document, second, they want to add annotations on basis of consultation of a lexicon, third, they want to jump to instances of lexicon entries in an annotated corpus, fourth, they want to jump to a lexicon entry from an annotation, fifth, they want to use a lexicon for the formulation or execution of queries on annotated corpora.

Finally, next to linguistic work, there is a large cultural and ethnological component to documentation of especially endangered languages. Good examples are the widely felt need for annotations of music, dance, rituals, etc in such a way that they can be inspected and analyzed in relation to linguistic annotations.

Gesture and sign language research

At MPI and the University of Nijmegen (UN) we are faced with the following studies that push the requirements for an efficient framework for manually creating multimodal annotations (only some will be mentioned here):

- Gesture studies where gestures in various contexts and from various cultural backgrounds are compared.
- Multimodal interaction studies where the precise timing between the speech and gesture modalities are analyzed to distinguish production models.
- Studies where the different types of gestures used in minority languages are analyzed (Enfield, 2002).
- Studies where the differences between several European sign languages are analyzed (Crasborn, 2003).
- Studies where the differences between sign languages world wide are analyzed (Zeshan, 2004).

In these types of studies often many different annotation layers are needed, for example to annotate different articulators. We have seen cases of up to 50 layers. These layers are either completely independent with respect to their time alignment, or they can be explicitly dependent.

Since there can be so many layers, often associated with controlled vocabularies, it is required that complete specifications for such tier setups can be made available in repositories for re-use.

Because gesture research and sign language annotation is mainly based on video recordings and because it is concerned with details on a very short time scale, there are high demands on video handling. Synchronized playback of multiple video recordings of the same event is necessary, MPEG2 support and video zooming are desirable, video frame accurate annotation is a necessity.

A highly desirable feature is the support for the annotation of spatial regions of the video signal during some time interval, for example to mark relevant locations or areas, or trajectories over time.

For gesture and sign language studies it is sometimes required that other types of media than video and audio can be visualized and used as the basis for annotation. Examples are eye tracking or data glove time series. It is necessary that each of those signals can be visualized using a time axis that is shared with audio and annotation data.

Collaborative annotation

A problem that all user groups share is that they want to collaborate on annotation projects from different geographical locations. ELAN is therefore in the process of being extended to support peer-to-peer cooperation. A group of users can share an annotation document, potentially including streamed video and audio data, during a working session. Users can chat, they can point at elements, times and locations in the document viewers and they can propose and commit changes to the document. All of this is propagated instantly to all participants using peer-to-peer technology[7].

This is more thoroughly discussed in (Brugman, Crasborn, Russel, 2004)

ELAN's main functions

For the design of ELAN a number of guiding principles are used:

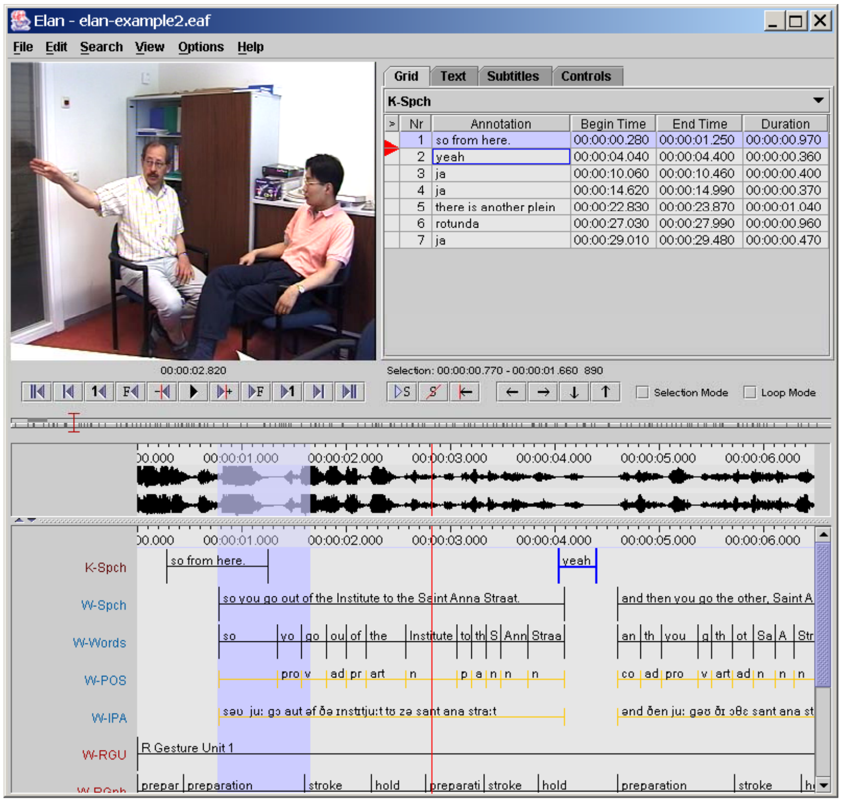
- As is common practice in software engineering, representations of annotation structures on the screen or on print are decoupled from representations used for persistence, document exchange or searching.

[5] http://www.volkswagen-stiftung.de

[6] Widely used in field linguistics. Typically blocks of text with parallel lines, where association of tokens across lines is represented by vertical text alignment

[7] Implementation is done using JXTA, http://www.jxta.org

- Several alternative viewers on the same underlying annotation data are supported. Each viewer is optimized to support certain tasks.
- All viewers are synchronized with respect to media time, selected time interval and active annotation. Modifications can be made in each viewer and show up in all other viewers instantaneously.


Elan - elan-example2.eaf
File Edit Search View Options Help
Grid Text Subtitles Controls
K-Spch
00:00:02.820
Selection: 00:00:00.770 - 00:00:01.660 890
Selection Mode
Loop Mode
00.000 00:00:01.000 00:00:02.000 00:00:03.000 00:00:04.000 00:00:05.000 00:00:06.000
K-Spch
so from here.
yeah
W-Spch
so you go out of the Institute to the Saint Anna Straat.
and then you go the other, Saint A
W-Words
so yo go ou of the Institute to th S Ann Straa
an th you g th ot Sa A Str
W-POS
pro v ad pr art n p a n n n
co ad pro v art ad n n n
W-IPA
səʊ ju: gɔ aut əf ðə ɪnstɪtju:t tʊ zə sant ana stra:t
ənd ðen ju: gəʊ ðɪ ɔθɛ sant ana st
W-RGU
R Gesture Unit 1
W-RGph
prepar preparation stroke hold preparati stroke hold preparation stroke h

>	Nr	Annotation	Begin Time	End Time	Duration
	1	so from here.	00:00:00.280	00:00:01.250	00:00:00.970
	2	yeah	00:00:04.040	00:00:04.400	00:00:00.360
	3	ja	00:00:10.060	00:00:10.460	00:00:00.400
	4	ja	00:00:14.620	00:00:14.990	00:00:00.370
	5	there is another plein	00:00:22.830	00:00:23.870	00:00:01.040
	6	rotunda	00:00:27.030	00:00:27.990	00:00:00.960
	7	ja	00:00:29.010	00:00:29.480	00:00:00.470

Figure 1: Screenshot of a document opened in ELAN 2.0

- We try to impose as few restrictions on user's annotation projects as possible. Numbers and types of annotation layers are therefore unrestricted and user definable. Time alignment on different tiers can be completely independent, or made dependent at the user's choice.
- As much information as possible is represented explicitly. We try to avoid implicit representation of annotation data such as codes that are embedded within annotation values, structure that is encoded by text alignment on a page or by the hierarchical structure of XML documents.
- We adhere to principles of stand-off annotation in the sense that different layers of annotation are kept separate (but not necessarily in different files).
- As much as possible, we adhere to standards, as for example Unicode.

Figure 1 shows a multi-layer annotation document opened in ELAN. ELAN, the displayed document and ELAN sources can be downloaded from the MPI tools website[8].

ELAN's document window shows several different panels or viewers, most of which are optional and can be detached as a separate window from the main window.

The upper left viewer shows the video signal making use of either the Java Media Framework (JMF) on Windows or QuickTime on Macintosh. When the video viewer is detached it can be scaled, for example to show the full resolution of MPEG-2. Two video viewers can be used to show two video signals that are recorded in sync. To the right of the video viewer a number of alternative annotation viewers can be made visible using tab panes. The Grid viewer shows a clickable list of annotations on a chosen tier with their begin and end times and durations, the Text viewer shows all annotation values on a chosen tier as running text. It is also clickable, editable, selectable and shows current media time. The Subtitle panel shows up to four selectable tiers as video subtitles that play along with media time. The control tab contains sliders of play back rate and audio volume.

The button panel shows groups of buttons for play/pause and stepping through time with several step sizes, for operating on the time selection and for jumping from annotation to annotation.

An annotation density viewer shows where annotations exist between the beginning and end of the document's media files. A wave form panel shows sample data for mono or stereo speech.

In the bottom panel two alternative annotation viewers can be shown: the Timeline viewer and the Interlinear viewer. The Timeline viewer shows annotations for each tier as a time segment with a text label. Black segments represent annotations that are or can be aligned with media time. Yellow segments represent annotations that refer to other annotations. Their begin and end times are derived from their parent annotation's begin and end times. Annotation tiers can be made visible or invisible, or can be reordered with a simple drag-and-drop operation. The Interlinear viewer (not shown) shows groups of hierarchically connected annotations as interlinear text.

New documents are created by selecting one or more media files and, optionally, defining their time origins. The next step is to define Linguistic Types for annotation tiers. Such a definition specifies the semantics of annotation values, whether annotations are time alignable or refer to other annotations, and which constraints hold on annotation values or on structural connections with other annotations (see chapter on ACM).

Then Tiers can be defined and associated with Linguistic Types. Tiers can be independent or be connected to a parent tier. Annotations can now be created on each tier by simple user operations, taking constraints into account.

ELAN's user interface can be localized on the fly by selecting a language from a menu.

[8] http://www.mpi.nl/tools

Abstract Corpus Model

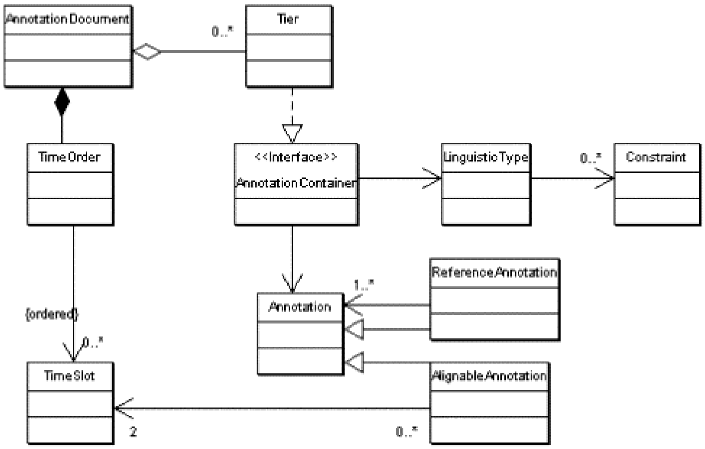


Figure 2: class diagram of the core part of the Abstract Corpus Model

Figure 2 shows the core part of the ACM. Tiers are containers for annotations. Annotations associated with these tiers have one of two reference types: (1) Annotations can be aligned with media time and then have a begin and end represented by a TimeSlot, or (2) they refer to one or several other annotations. TimeSlots can be explicitly aligned with media time but do not have to be. However, all TimeSlots are explicitly ordered within the AnnotationDocument by means of a TimeOrder.

Finally, Tiers are associated with LinguisticTypes that can in turn be associated with Constraints. By implementing stereotypic sets of constraints in program code annotations can be connected in complex patterns. A number of these stereotypes are already implemented:

- Time subdivision: annotations on a dependent tier are all within the time interval of an annotation on the parent tier, and between annotations with the same parent tiers no time gaps are allowed (example: gestures can be decomposed into separate gesture phases).
- Symbolic subdivision: annotations on a dependent tier refer to annotations on a parent tier. Annotations that point to the same parent annotation are explicitly ordered (example: words can be decomposed into morphemes).
- Symbolic association: there is a 1-1 relation between a dependent annotation and it's parent annotation (example: all cases where annotations can have some attribute value, like part-of-speech on a word or morpheme)

A few other stereotypes are planned to be implemented:

- Annotations on a dependent tier can refer to one or more annotations on a specific parent tier. These parent annotations do not have to be consecutive. This stereotype can be used to model for example co-reference chains.
- Dependent annotations can refer to one or more annotations on the same tier and on a specific parent tier. This makes recursive trees, like syntax trees, representable.

Using these basic elements and stereotypes it is possible to represent very complex annotation documents. In the area of Endangered Languages it would for example be possible to combine time aligned phonetic transcriptions for several speakers with interlinear text analysis, and with gesture and musical annotation. Comments could be attached to combinations of annotations on any of these tiers.

Conclusion

Although the growth of ELAN's user base confronted us with diverging requirements, careful modeling and using a set of proven design principles for annotation tools helped us cope with that. ELAN development is now in a state that allows straightforward expansion to cover new user needs.

Moreover, the latest developments and insights on annotation tools, formats and standards seem to converge. Work on ELAN is consistent with this convergence, and we hope that it is actually contributing to it.

References

S. Bird and M. Liberman. 2001. *A formal framework for linguistic annotation.* Speech Communication, 33(1,2):23-60.

S. Bird and G. Simons. 2003. *Seven dimensions of portability for language documentation and description.* In Bojan Petek (ed.), Portability issues in human language technologies: LREC 2002

H. Brugman and P. Wittenburg. 2001. *The application of annotation models for the construction of databases and tools.* IRCS Workshop on Linguistic Databases, University of Pennsylvania. Philadelphia.

H. Brugman, O. Crasborn and A. Russel. 2004. *Collaborative annotation of sign language data with peer-to-peer technology.* To be published. LREC 2004.

O. Crasborn. 2003. *Internal ECHO report* (to be published). Nijmegen.

S. Drude and H. Lieb. 2002. *Advanced Glossing – a language documentation format and its implementation with Shoebox.* International Workshop in Field Linguistics at LREC 2002. Las Palmas.

N. Enfield. 2002. *Hand pointing in Laos: form and function in a locality description task.* MPI Annual Report 2002. Nijmegen.

B. MacWhinney. CHILDES. http://childes.psy.cmu.edu

U. Zeshan. 2004. *Sign Language Typology Project* (to be published). Nijmegen.