The Eclipse Annotator: an extensible system for multimodal corpus creation

Fabian Behrens, Jan-Torsten Milde

Department of Computer Science, Fulda University of Applied Sciences
fabian.behrens@informatik.fh-fulda.de, milde@fh-fulda.de

Abstract

The Eclipse-Annotator is an extensible tool for the creation of multimodal language resources. It is based on the TASX-Annotator, which has been refactored in order to fit into the plugin based architecture of the new application.

1. Introduction

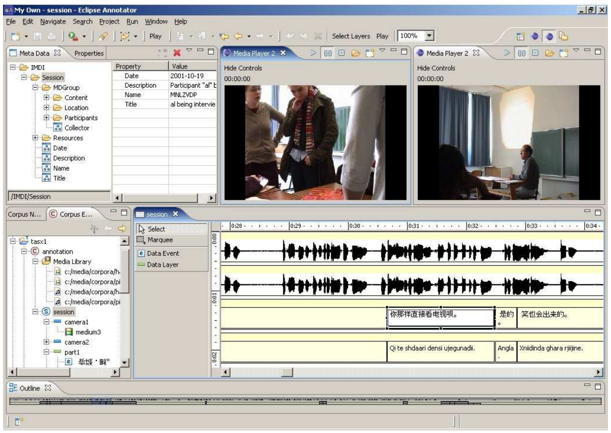
In this paper we will show how an existing linguistic tool, the TASX-annotator, has been transformed into a set of Eclipse plugins which define an extensible system for the creation of multimodal language resources: the Eclipse-annotator.

The robust extensible software infrastructure provided by Eclipse (see `www.eclipse.org`) makes it possible to efficiently design and implement linguistic applications. Eclipse provides excellent components (e.g. configurable editors, tree views) which are perfectly matching the structures of the scientific data under investigation. Existing tools are easily integrated into an Eclipse application. Furthermore, Eclipse-based tools are well portable to a number of operating systems.

2. From TASX-Annotator

Today, the design and implementation of most linguistic tools is likely to be driven by the specific interests of a small group of scientist. The TASX-Annotator is a good example of such a linguistic tool ((Milde and Gut, 2001), (Milde and Gut, 2002a), (Milde and Gut, 2002b), (Milde and Gut, 2003),(Schonefeld and Milde, 2004). It started out as a small sized project with a simple function and eventually turned into an unmanageable complex system integrating a large number of Java libraries controlled by a self defined component framework.

The system has shown to be useful in setting up an number of multimodal corpora. Within these projects, the TASX-Annotator has helped to integrate data from different sources. It allowed to transcribe video streams and convert the results into various formats needed for the analysis. Still, of the four stages of the corpus creation process (*design*, *collection*, *annotation and analysis*), the annotation phase was best supported by the tool. A large number of XSL-T scripts and additional tools were integrated or even had to be developed during the different phases. As a result, setting up a large corpus turns out to be inefficient and error prone. Even worse, until now, no suitable system is available to efficiently query large XML annotated corpora. To the best of my knowledge, the formentioned critical points hold true for many of the currently available tools.



Property	Value
Date	2001-10-19
Description	Participant "al" b
Name	MNLZVDP
Title	al being intervie

你那样直接看电视呗。 是的。 笑也会出来的。

Qi te shdaari densi ujegunadii. Angla. Xniidinda ghara rjiijine.

Figure 1: Eclipse-annotator perspective: multiple viewers (audio, video, corpus browser) and editors (partiture, metadata) are combined to form the main perspective.

3. To Eclipse-Annotator

The Eclipse-annotator[1] preserves the basic functionality of the TASX-Annotator. The user is able to annotate audio and video files on multiple levels. All data is stored in the TASX format, a generic XML based exchange format. A number of new modules have been implemented, making the new system more powerful:

- The underlying TASX data model has been extended in order represent a complete corpus including external media files.
- The software has been ported to the current Java version. Most classes have been refactored. A large number of unit tests have been defined to ensure system integrity during the migration process.
- The system is able to load a complete corpus, not just a single recording. The powerful corpus browser allows to create and manage very large corpora.
- It is possible to visually annotate the video by marking the frames graphically.
- The system is able to automatically identify hard cuts in the videos, making the handling of large video more efficient.

[1] The Eclipse-Annotator project is hosted at `sourceforge.net` under the name of `eclipse-annotat`.

- An XQuery engine has been integrated.
- It is now possible to use the platform specific video and audio software, thereby increasing the performance substantially.

The migration process has increased the functionality, the stability, the usability and the reliability of the tool.

4. The TASX format

All linguistic data is stored in an XML-based format called TASX: the *T*ime *A*ligned *S*ignal data e*X*change format. With TASX it becomes possible to to transform, query and distribute the content of multimodal corpora, and to perform adequate linguistic analysis (Milde and Gut, 2002b).
A TASX-annotated corpus consists of a set of *sessions*. A session basically represents a single *recording* or *experiment* of a multimodal corpus. Each session is holding an arbitrary number of descriptive tiers, called *layers*. The layers of a session are treated as time aligned sets, which are directly or indirectly linked to the primary data. Layers are used to distribute the annotations of a session and thus simplifying the information access. Each layer consists of a set of independent *events*. The events store textual information (e.g. a syllable or a handform) and are linked to the primary audio data by two time stamps denoting the interval of this event. Events form the annotations atoms of a TASX annotated corpus. With respect to the formal definition of TASX, events are the leaves of the XML tree. No further formal restriction is imposed on the content of an event.
In order to manipulate arbitrary media content, a media library has been introduced as a child of the top level element tasx. Again no restrictions are imposed onto the linkable file type. This makes it possible to add video and audio file as well as OpenOffice Documents or PDF files.
Sessions, layers and events all carry linking attributes. All elements of a TASX annotated corpus must provide a uniqe ID (*s-id*, *l-id* and *e-id*)). In addition cross-references between the elements can be specified using the optional (*ref*) attribute. A validating XML parser should be able to check the existence of the IDs and, if cross references are provided, the consistincy of the intra document linking. The ID/IREFS mechanism represents a means to further restrict the *content* of a TASX annotated corpus, e.g. a session *ref*-attribute could register all IDs of the enclosed layers. The following DTD fragment formalizes the TASX format:

```
<!ELEMENT tasx (meta*, media?, session+)>

<!-- the media library -->
<!ELEMENT media (medium+)>
<!ELEMENT medium (desc*)>

<!ELEMENT session ((meta*, layer)*)>

<!ELEMENT layer ((meta*, event)*)>
<!ELEMENT event (#PCDATA|meta)*>

<!ELEMENT meta (desc*)>
<!ELEMENT desc (name,val)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT val  (#PCDATA)>

<!-- ATTRBUTES -->
<!ATTLIST meta
         m-id CDATA #REQUIRED
         access CDATA #IMPLIED
         level CDATA #IMPLIED
>

<!ATTLIST medium
         md-id CDATA #REQUIRED
>

<!ATTLIST session
         s-id CDATA #REQUIRED
         day CDATA #REQUIRED
         month CDATA #REQUIRED
         year CDATA #REQUIRED
>

<!ATTLIST layer
         l-id CDATA #REQUIRED
>

<!ATTLIST event
         e-id CDATA #REQUIRED
         start CDATA #REQUIRED
         end CDATA #REQUIRED
         mid CDATA #IMPLIED
         len CDATA #IMPLIED
>
```

4.1. TASX-level 1

A TASX-annotated corpus, that *directly* links the primary data and the corpus data by defining a temporal interval in the start/end attributes of an event, is called a *TASX-level 1* corpus. TASX-level 1 provides a solution for one of the most common problems of XML-annotated multimodal corpora: the temporal *overlap* of annotation units. XML files define a tree structure and as such an overlap of opening- and closing tags is not allowed. This limitation is too restrictive for a large number of linguistic application areas. Within the TASX-annotated XML file the events are ordered linearly, but their scope is defined in the interval encoded in the attributes. Accordingly events may overlap in time. This mechanism can also be applied to primary data, which has no intrinsic temporal order, respectively to data, where the temporal order is lost (e.g. it is complex to reconstruct the temporal order of SyncWriter files, because the segmentation lists do not carry any temporal information). In order to describe temporal overlap in such primary data, a set of linearly ordered reference points has to be defined. This can be achieved by simply refering to the set of natural numbers.

4.2. TASX-level 2

A TASX corpus, which extends the direct temporal linking of events to *linking events to other events* is called a *TASX-level 2* corpus. TASX-level 2 allows to define hierarchical relations between annotation layers. These inter layer relations can be established, because the formal description of TASX does not restrict the values of the start/end attributes in any way. Therefore arbitrary strings can be used to describe the relations to other layers and events on these layers. These strings might contain XPointer or XPath expressions (Wilde and Lowe, 2002) or being formulated in any other suitable syntax. With respect to linguistic research, this TASX approach allows to represent cross level relations, e.g to denote the hierarchical relation between words and syllables. In a case study in conjunction with Voormann (Voormann et al., May 2004) hierarchical relations between different layers of the LeaP corpus have been computationally constructed. Implementations in related tools follow the same approach, e.g. the Elan annotation tool designed by Brugman und Wittenburg (Brugman and Wittenburg, 2001) provides a mechanism to constrain the relations between annotation layers.

TASX-level 1 and 2 are comparable to standard approaches currently discussed in the field of computational corpus linugistic (Bird and Liberman, 1999). Schmidt provides Exmaralda, a tool for conversational analysis, which follows the Bird & Liberman approach (Schmidt, 2001). Alternatively stand-off markup is proposed by MATE (Dybkjaer et al., June 1999), respectivly NITE (Carletta et al., 2002). , while Kipp (Kipp, 2001) specifies hierarchical relations between tiers.

4.3. TASX-level 3

TASX defines a *data centric* information model. It focuses on the organisation of the data and leaves the internal logical structure of the primary data untouched. The TASX model basically organizes the data of a session as a two dimensional array. A row of the array is equivalent to an annotation layer, each field of the array is mapped to an event. The content of the fields is unrestricted.

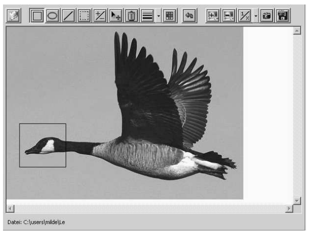


Figure 2: Eclipse-annotator plugins: the image annotator displays snapshots of the currently loaded video. The frames can be visually annotated. The annotation is stored in an SVG conformant format.

The clear advantage of such an approach is the fact, that arbitrary event descriptions can be gathered in a TASX conformant corpus. Within in the context of the LeaP corpus we were able to encode orthographical representations of words and syllables, but also phonemes encoded as SAMPA strings and phrasal tone gradients encoded according to the ToBI standard (Beckman and Elam., 1994).

It is also possible to include primary data with a more complex structure, as long as the data can be serialized and represented as a linear string. One possible solution is to encode the data as Base64 ASCII strings. This approach is used inside the TASX-annotator to store multi-line comments entered by the user. Finally XML-annotated strings could be stored as the content of an event. In this case, the special characters of XML have to be encoded with their entity counterpart.

It is due to this flexibility, that TASX is able to integrate a large number of currently available representational formats for linguistic data without information loss. At the same time this also marks a central problem of the data centric approach. The internal structure of the data stored in an event is completly transparent to TASX. As a consequece, there is no direct way to ensure the consistency and validity of the data stored (at least not with the formal definition of TASX being descriobed as a DTD).

As such TASX seemingly dismisses the two most important advantages of XML: the *structure guided creation* of annotated content, based on the formal description in form of a DTD and the *automatic verification of the linguistic structure* of a corpus using a standard validating XML parser.

With *TASX-level 3* we try to define a balanced compromise between the data centric and the structure centric approach. The approach tries to decouple the XML elements of the underlying data centric TASX format from the XML elements describing the semi structured linugistic data. To achieve this, a TASX namespace has been defined. While this is not possible with DTDs, we used XML schema to formally define the namespace. An XML parser processing a TASX corpus will be able to distinguish between the TASX elements and the embedded elements. The approach thus allows to encode tree like structures with XML and stores them as part of a TASX corpus. The embedded XML structures must be wellformed. TASX-level 3 therefore allows to setup corpora that e.g. combine syntactic trees and phonetic transcriptions. References between the different layer can be established.

Despite of its simplicity, the TASX-format is powerful enough to encode most of the corpus annotation formats currently in use. Indeed a number of format transformation programs have been implemented. In order to e.g. reconstruct the equivalent annotation graphs (Bird and Liberman, 1999) representation of a TASX annotated corpus, one only has to collect the time stamps encoded in the start and end attributes of the event tags, sort them and then produce the timeline. Finally the time stamps of the events have to be replaced by references to the timeline.

While the TASX-Annotator has not been capable of handling anything beyond TASX Level 1, the Eclipse Annotator is able to handle TASX Level 2. A number of predefined constraints have been implemented. These include the temporal relations defined by Allen and crosslevel containment relations. Until now, the constraints have been implemented in Java directly. This is likely to change in the future.

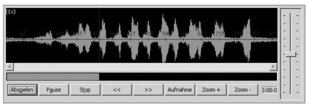


Figure 3: Eclipse-annotator plugins: the audio player displays the oszillogram of the currently loaded sound file.

5. System architecture

The Eclipse workbench provides a plugin mechanism, allowing users to extend and modify its functionality. When adopting to this mechanism, linguistic applications are able to use the Eclipse software infrastructure. In addition, Eclipse provides the Rich Client Platform (RCP), making it simple to create standalone Eclipse based applications.

The Eclipse-Annotator has been distributed onto a number of independent plugins. These plugins implement editors and viewers for the underlying TASX-based data model. In addition a set of multimedia views have been developed to display the primary data. GUI and applications logic are strictly seperated following the MVC (model view controller) design pattern. In order to communicate with each other, the plugin register themselves as event listeners or event providers with the model. This kind of inter process commnication is supported by the Eclipse framework. The application will run without errors, even if a view is not activated. All implemented plugins can thus be reused in other Eclipse based applications. The following plugins are available:

1. The TASX data model: this plugin implements the extended TASX data model and provides interface function to create and modify the model.

2. GUI: a set of viewers for the data model. A central component is the corpus browser view. It allows to load a corpus incrementally and displays its content using a hierarchical tree view (see figure 1).

3. Session view: this view is implemented using GEF, the Graphic Modeling Framework. It displays a session as an partiture. Each event is displayed as a rectangle. The text can edited in place.

4. Video player: two versions of the video player exist. One is based on the JMF, the other uses the Microsoft video playback engine (see figure 1).

5. Audio player: the audio player displays the oszillogram of the loaded sound. It is possible to zoom in and out (see figure 3).

6. Image view: this view displays screenshots of the video. It is possible to visually annotate the content of the image. Visual annotation is stored in an SVG conformant format(see figure 2).

6. Conclusion

The Eclipse-Annotator is a versatile tool for the creation of richly annotated corpora. By splitting up the application into a set of independent Eclipse plugins a high reusablity of the components has been achieved. The open architecture will simplify the integration of new functionality.

7. References

Mary E. Beckman and Gayle Ayers Elam. 1994. Guidelines for ToBI Labelling. Technical report, Ohio State University. Version 3.0, March 1997.

S. Bird and M. Liberman. 1999. A Formal Framework for Linguistic Annotation. Technical Report MS-CIS-99-01, Department of Computer and Information Science, University of Pennsylvania.

Hennie Brugman and Peter Wittenburg. 2001. Mpi tools for linguistic annotation. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

J. Carletta, D. McKelvie, and Isard A. 2002. Supporting linguistic annotation using xml and stylesheets. In G. Sampson and D. McCarthy, editors, *Readings in Corpus Linguistic, Continuum International*.

L. Dybkjaer, M. B. Moeller, N. O. Bernsen, J. Carletta, A. Isard, M. Klein, D. McKelvie, and A. Mengel. June 1999. The mate workbench. In David Traum, editor, *Proceedings of ACL'99, Demonstration Abstracts. University of Maryland*, pages 12 – 13.

Michael Kipp. 2001. Anvil - a generic annotation tool for multimodal dialogue. In *Proceedings of the Eurospeech 2001, Aalborg*, pages 1367 – 1370.

J.-T. Milde and U. B. Gut. 2001. The TASX-engine: an XML-based corpus database for time aligned language data. In Peter Buneman Steven Bird and Mark Liberman, editors, *IRCS Workshop on Linguistic Databases, University of Pennsylvania, Philadelphia, USA*.

J.-T. Milde and U. B. Gut. 2002a. A prosodic corpus of non-native speech. In B. Bel and I. Marlien, editors, *Proceedings of the Speech Prosody 2002 conference, 11-13 April 2002. Aix-en-Provence: Laboratoire Parole et Langage*, pages 503 – 506.

J.-T. Milde and U. B. Gut. 2002b. The tasx-environment: an xml-based toolset for time aligned speech corpora. In *Proceedings of the third international conference on language resources and evaluation (LREC 2002, Gran Canaria*.

Jan-Torsten Milde and Ulrike Gut. 2003. Annotation of Conversational Gestures using TASX and CoGesT. *KI - Gesellschaft fr Informatik, special issue on Embodied Conversational Agents*.

T. Schmidt. 2001. Gesprächstranskription auf dem Computer - das System EXMARaLDA. *Gesprächsforschung, http://www.gespraechsforschung-ozs.de*, 2.

Oliver Schonefeld and Jan-Torsten Milde. 2004. Embedding imdi metadata into a large phonetic corpus. In *LREC 2004, Forth International Conference on Language resources and Evaluation, Lisbon, Portugal*.

Holger Voormann, Ulrich Heid, Jan-Torsten Milde, Ulrike Gut, Katrin Erk, and Sebastian Pado. May 2004. Flexible querying of xml-encoded multi-layer corpora. In *Proceedings of LREC 2004, Lisbon*.

Erik Wilde and David Lowe. 2002. *XPath, XLink, XPointer, and XML*. Addison-Wesley Professional.