ANNEX – a web-based Framework for Exploiting Annotated Media Resources

Peter Berck, Albert Russel

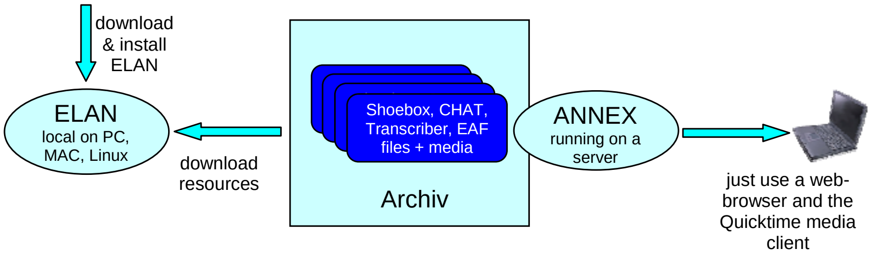
Max-Planck-Institute for Psycholinguistics
Wundtlaan 1, 6525 XD Nijmegen, The Netherlands
peter.berck@mpi.nl

Abstract

Manual annotation of various media streams, time series data and also text sequences is still a very time consuming work that has to be carried out in many areas of linguistics and beyond. Based on many theoretical discussions and practical experiences professional tools have been deployed such as ELAN that support the researcher in his/her work. Most of these annotation tools operate on local computers. However, since more and more language resources are stored in web-accessible archives, researchers want to take profit from the new possibilities. ANNEX was developed to fill this gap, since it allows web-based analysis of complex annotated media streams, i.e., the users don't have to download resources and don't have to download and install programs. By simply using a normal web-browser they can start their linguistic work. Yet, due to the architecture of the Internet, ANNEX does not offer the options to create annotations, but this feature will come. However, users have to be aware of the fact that media streaming does not offer that high accuracy as on local computers.

1. Introduction

Much of linguistic work has to do with creating and analyzing annotations of texts, sound and/or video files either manually as in field linguistics and multimodal research or automatically as in Natural Language Processing. The generated structures can become rather complex as was described for example by Liberman&Bird [1], Brugman&Wittenburg [2], the NITE project [3] or within the Tipster initiative [4]. The individual annotations are associated with annotation types – often also called tiers – and they have relations in time and linguistic structure. Some tiers include special encodings such as representations of syntactic structure for example in the TIGER format [5]. We distinguish these kinds of structural annotations from semantic annotations where singular assertions are made on content elements and which are typically encoded for example with the help of RDF triples [6].



This figure shows the relation between ELAN and ANNEX.

These structural annotations encode linguistic knowledge and are important for the documentation of linguistic phenomena and for further linguistic analysis. For researchers it is therefore of greatest importance to have efficient and convenient tools to create, analyze and visualize annotations together with their original streams – be it video, audio or textual streams. A large set of text annotation tools and frameworks has been developed for NLP type of applications such as GATE [7] as well as for field and corpus linguistics such as Shoebox [8] and CLAN [9]. Also some tools are known that allow users to annotate sound files such as Transcriber [10]. Only a few tools such as ELAN [11] are known that support all three streams (text, audio and video). All these tools differ in efficiency and convenience, functionality and expressive power of the underlying data model.

Almost all of the better tools are operating as local tools. However, at the MPI for Psycholinguistics we see an increasing need to provide simple to use web-based frameworks to allow users to utilize language archive content without having to first install a local tool and then download a resource bundle that can easily require more than 1 GB of disc space and an appropriate download time. Before starting with an in-depth analysis or further processing some users want to carry out quick inspection of the complex annotations together with the annotated streams. This is the reason that the MPI developed the web-based ANNEX framework [12].

2. Usage Scenarios

At the MPI we have collected a large repository covering annotated media files from many different projects and individual researchers. The contributions come from projects documenting endangered languages (DOBES [13]), language acquisition studies (ESF [14]), sign-language studies (ECHO [15]), bilingualism projects (DBD [16]), gesture studies (Enfield [17]), gathering the Dutch National Spoken Corpus (CGN [18]) and many others. This implies that they are created with the help of different tools such as already mentioned which produce a variety of output formats, that the annotation structures are different both with respect to the amount of tiers and their complexity and of course the nature of the linguistic encoding. Here we will ignore the differences in character encoding although they partly form big problems.

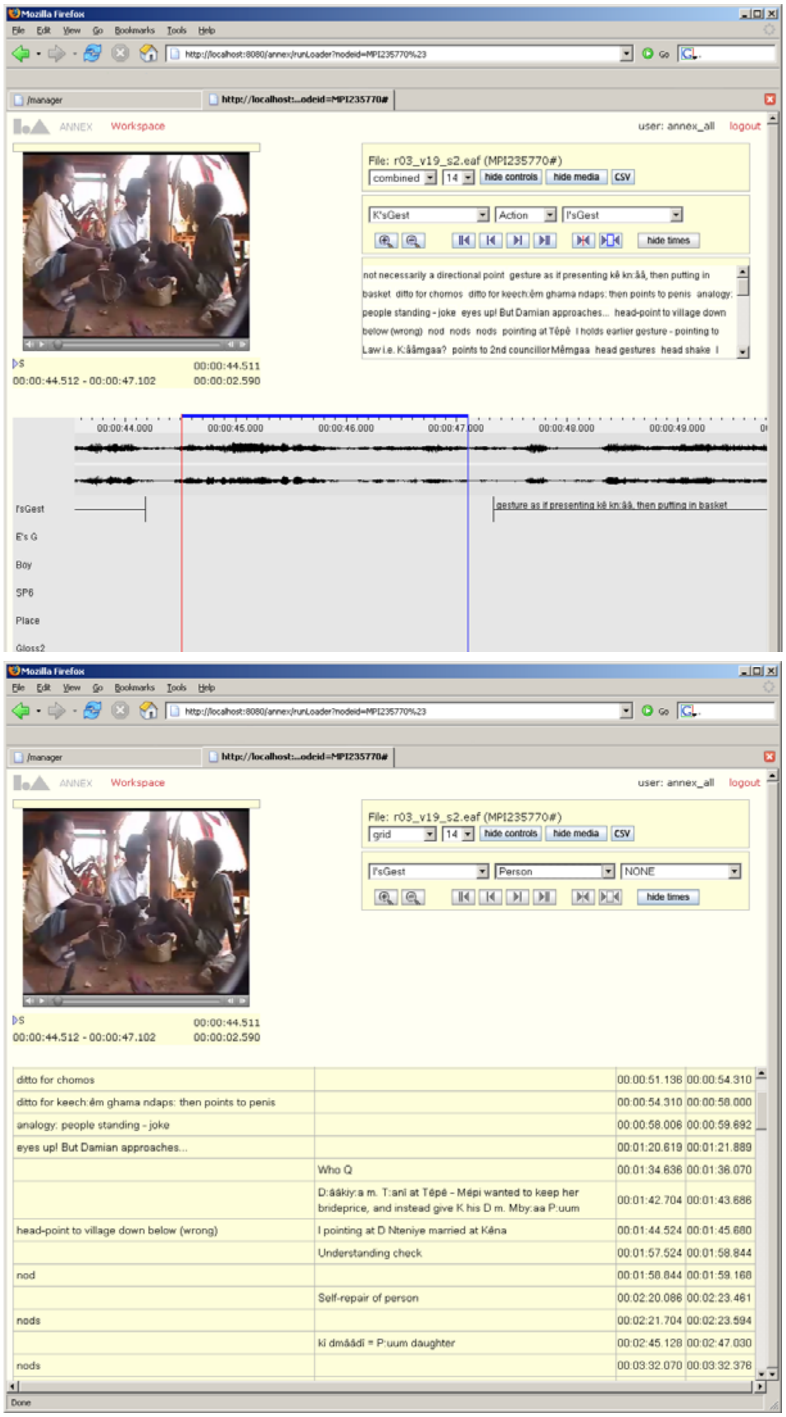
With the exception of some projects these differences occur as well where there is an interest from researchers to overcome the barriers. We understood that researchers want, for example, to compare utterances from languages that are spoken in geographically adjacent regions to find cognate sets and mutual influences. This requires a framework which allows users to select a number of resources with the help of metadata browsing or searching, to formulate queries about annotation contents and to visualize the material indicated as hits. Therefore, the framework has to overcome the differences in annotation structures and formats to a certain extent and has to offer some means to overcome the differences in linguistic encoding.

Many other usage scenarios can be thought of which will require a web-based framework. Members of language communities, for example, would like to access annotated media streams directly by selecting a resource via geographic browsing. Sign language speakers would like to compare two signs immediately via the web etc. In the outreach section we will mention a few additional scenarios that have to be supported.

3. ANNEX

ANNEX is meant to fill this gap. It is a server-based tool that allows users to access annotated media streams via the web. It can be operated by using a standard HTML browser and it makes use of HTTP transfers except for video. Video streams are served by the Darwin video streaming server that makes use of the RTSP protocol [19] and requires the Quicktime client player. In doing so only the selected video fragment has to be downloaded and visualized. For video streaming an MPEG4 codec is used, i.e., for all video stream in the archive an MPEG4 copy has to be created.

In many respects ANNEX is comparable to ELAN. The screenshots may give an indication of a typical layout which can be defined by the user. The annotations can be visualized in different styles dependent on the goals of the researcher. The tiers to be presented can be selected and their order can be



S
00:00:44.512 - 00:00:47.102
00:00:44.511
00:00:02.590

not necessarily a directional point gesture as if presenting kê kn:åå, then putting in basket ditto for chomos ditto for keech:êm ghama ndaps: then points to penis analogy: people standing - joke eyes up! But Damian approaches... head-point to village down below (wrong) nod nods nods pointing at Têpê I holds earlier gesture - pointing to Law i.e. K:ååmgaa? points to 2nd councillor Mêmgaa head gestures head shake I

gesture as if presenting kê kn:åå, then putting in basket

S
00:00:44.512 - 00:00:47.102
00:00:44.511
00:00:02.590

ditto for chomos		00:00:51.136	00:00:54.310
ditto for keech:êm ghama ndaps: then points to penis		00:00:54.310	00:00:58.000
analogy: people standing - joke		00:00:58.006	00:00:59.692
eyes up! But Damian approaches...		00:01:20.619	00:01:21.889
	Who Q	00:01:34.636	00:01:36.070
	D:ååkiy:a m. T:anî at Tépê - Mépi wanted to keep her brideprice, and instead give K his D m. Mby:aa P:uum	00:01:42.704	00:01:43.686
head-point to village down below (wrong)	I pointing at D Nteniye married at Kêna	00:01:44.524	00:01:45.680
	Understanding check	00:01:57.524	00:01:58.844
nod		00:01:58.844	00:01:59.168
	Self-repair of person	00:02:20.086	00:02:23.461
nods		00:02:21.704	00:02:23.594
	ki dmåådî = P:uum daughter	00:02:45.128	00:02:47.030
nods		00:03:32.070	00:03:32.376

The figure above shows a video scene, a time-line viewer with sound information, linked annotation information, the annotations of a selected tier and the usual controls. The figure below offers the same scene, however, the text is presented in a grid type of style. All elements can be rearranged on the screen with the help of mouse operations.

determined by the user.

Synchronized viewers, fragment selection based on annotations or time period marking, the capability to cope with a wide range of annotation structures and the support for a number of formats (EAF, Shoebox, CHAT, CGN – for Transcriber files a converter is available) make ANNEX a useful tool to access annotated media streams. For all access attempts – even when searching – the access permissions are checked. Only when access permissions are given a certain resource can be visualized or subject of searching.

4. Searching and Indexes

Another very useful feature is the search component. The user can first execute metadata queries on the whole or a sub-archive. The resulted set of resources can be subject of unstructured or structured searches. In unstructured searches the specified pattern is matched against all annotations found in the selected resources. In structured searches the user can associate patterns with tiers and combine them over time to complex queries. Parsers for the different formats mentioned above are available so that the search operation can include CHAT, Shoebox and EAF files. The hits can be selected, be started and directly be analyzed with ANNEX. Operating in a cross-corpora scenarios raises the problem of interoperability at a linguistic encoding level. Tier names and the encodings used on the tiers will differ. First simple methods are available to allow the user to define and store patterns across different corpora. However, to improve efficiency the inclusion of ontology mechanisms is required.

Searching is made fast by using index files. ANNEX will mainly be used in combination with archives. Therefore, it made sense to provide an API so that whenever a archive management and upload system will upload a new resource into the archive the index will automatically be upgraded. In this way LAMUS (the Language Archive Management and Upload System developed at the MPI [20] can interact with ANNEX. However, an interested user could use his own management system.

5. Geographic Browsing and Lexicon Interaction

Recently we started to make use of the geographic paradigm which can be a very strong one for accessing language resources. It was shown that it is very simple to start ANNEX from Google Earth, for example, and to directly visualize the selected resources.

Another first step was done in integrating annotation resources and lexica. When selecting a string in an annotation with the help of ANNEX the LEXUS lexicon tool can be started to show the corresponding entry. Here APIs are used that can also be used to implement more advanced interaction operations.

6. Next Steps

ANNEX was primarily designed as a visualization framework for two reasons: (1) With ELAN we can offer a professional tool to create annotations and (2) media streaming via the web does not offer the timing accuracy that is needed in many cases such as in multimodal interaction. However, first demonstrations to users revealed that users want to use ANNEX as well to create new or modify existing annotations. One of the next steps will be to extend ANNEX in this direction. This will have implications for managing the archive that will store the updates. For the integration of updates ANNEX will make use of an API provided by the LAMUS tool. It is LAMUS that has to check consistence, to create a new version of the modified resource, to create a new unique resource identifier entry etc.

It is obvious that we have to provide better mechanisms for cross-corpora operations. Currently, we are busy to create an editor that allows to create concept definitions and to define relations amongst them in a bottom-up driven fashion. This editor will also allow to store and exchange these knowledge resources and to link them with central ontologies such as being developed by ISO TC37/SC4 [20] and others. When ANNEX will be extended to allow to create new tiers and annotations it will be connected with the ISO TC37/SC4 and Shoebox concept registries via APIs that have already been successfully implemented in the LEXUS tool.

Another dimension of work will be to extend the interaction between corpora and lexica to allow more advanced operations. Here many functions can be thought of in both directions. In addition, ANNEX will be extended to also visualize TIGER compliant syntax trees in a time synchronous fashion. Due to the similarity with the ELAN code ANNEX will take advantage of ELAN extensions and vice versa.

7. Conclusions

With the development of the web-based ANNEX tool we have created new opportunities of accessing annotated media streams that are stored in a language resource archive. A large group of users likes to access archive content by making use of the normal web-browsers. Until now ANNEX only supports the analysis and visualization of such annotated media streams due to the limitations in timing accuracy. However, ANNEX will be extended to allow the creation or modification of annotations. For quite some time these will not be so accurate as it can be done with ELAN.

We see the great power of web-based applications for the future, in particular with respect to interlinking ANNEX with all sorts of other applications such as LEXUS, Google Earth etc. Therefore, we will invest time to enhance the functionality of ANNEX in the future.

8. References

[1] S. Bird, M. Liberman (2001): *A Formal Framework for Linguistic Annotation. Speech Communication* 33 (1,2), pp 23-60,
[2] H. Brugman, P. Wittenburg (2001): *www.ldc.upenn.edu/annotation/database/papers/Brugman_Wittenburg/20.2.brugman.pdf*
[3] www.ltg.ed.ac.uk/NITE
[4] http://www.cs.nyu.edu/cs/faculty/grishman/tipster.html
[5] http://www.ims.uni-stuttgart.de/projekte/TIGER/paper/
[6] http://www.w3.org/RDF/
[7] www.gate.ac.uk
[8] www.sil.org/computing/shoebox
[9] childes.psy.cmu.edu
[10] http://trans.sourceforge.net/en/
[11] www.mpi.nl/tools
[12] www.mpi.nl/annex
[13] www.mpi.nl/DOBES
[14] http://corpus1.mpi.nl/ds/imdi_browser/ >ESF
[15] http://corpus1.mpi.nl/ds/imdi_browser/ >ECHO
[16] http://corpus1.mpi.nl/ds/imdi_browser/ >DBD
[17] http://corpus1.mpi.nl/ds/imdi_browser/ >Enfield
[18] http://corpus1.mpi.nl/ds/imdi_browser/ >CGN
[19] http://www.cs.columbia.edu/~hgs/rtsp/
[20] http://www.tc37sc4.org/