Linking annotations
Steps towards tool-chaining in Language Documentation

Dorothee Beermann, Pavel Mihaylov, Han Sloetjes

Norwegian University of Science and Technology (NTNU), Trondheim, Norway
Ontotext, Sofia, Bulgaria
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
dorothee.beermann@hf.ntnu.no, bin@bash.info, han.sloetjes@mpi.nl

Abstract

The strong point of a virtual research environment (VRE) is that it facilitates a collaborative approach to data management and creation. Ideally, such a VRE allows its users to combine best-for-the-task tools in a simple but efficient manner. This paper presents work in progress which aims to improve combined system use in Language Documentation. Exploiting the strength of the multi-media annotation tool ELAN and the online multi-lingual database TypeCraft, we describe how the annotation of audio, video and text resources can be enhanced by presenting technology which supports shared annotations and collaborative editing online. We examine data mobility problems by presenting a use-case, and discuss how a workflow typical for data processing in Language Documentation can be improved substantially through minor but user-oriented development.

Keywords: virtual research environment, language documentation, user-driven software development

1. Introduction

The strong point of a virtual research environment (VRE) is that it facilitates a collaborative approach to data management and creation. Ideally, such a VRE allows its users to combine best-for-the-task tools in a simple but efficient manner. This paper presents work in progress which aims to improve combined system use in Language Documentation. Exploiting the strength of the multi-media annotation tool ELAN and the online multi-lingual database TypeCraft, we describe how the annotation of audio, video and text resources can be enhanced by a technology which supports shared annotations and collaborative editing online. We examine data mobility problems encountered in this technology. The combined used of TypeCraft and ELAN is already practised, but it comes at present at a high cost. We present the technical changes necessary to improve the workflow for a combined system use. Some of the solutions presented here are already implemented while others still need to be finished.

2. Short system descriptions

2.1 TypeCraft

TypeCraft (TC) is an online database featuring a tabular editor for the manual creation of Interlinear Glossed Text (IGT). The core application is wrapped into a customised mediawiki (TC-wiki) which serves as a general entrance port and collaboration tool. Below we provide an overview of the application's main functionalities:

Annotation
- Manual import of continuous text or sentence collections
- Tabular interface for morpheme level glossing, automatic sentence break-up
- Drop-down reference lists for annotation and flexible insertion and deletion of words and morphs
- Semi-automatic annotation and easy access to relevant information such as gloss definitions and an ontology of grammatical concepts from the annotation interface

Collaboration
- Graded access, individual work spaces
- TC texts and phrases have their own URI and thus can be acquired and exchanged freely online
- User groups can share data
- Collaborative editing of TC-wiki pages

Data export
- Export of annotated data to Microsoft Word, Open Office and LaTEX for paper publications
- Print-friendly versions of the TC web pages including exported database material
- Export of XML for automatic data processing

TC search is another strong point of the service that can be highlighted here. It allows complex searches on several tiers so that word and morpheme queries can be freely combined with a search for specific glosses or combinations of glosses, co-occurring either in a phrase, on a word, or on a morpheme. Search is graded, and can target the user's own data, as opposed to group data where search serves to establish inter-annotator consistency.

2.2 ELAN

ELAN is a standalone tool for the manual annotation of digital audio and video recordings. It offers generic media annotation functionality and most of its current features are not specifically designed for research in linguistics, but are useful in other areas of research as well. The same holds for its data model, with tiers as

containers for annotations and the possibility of defining hierarchical relations between tiers and, ipso facto, between annotations. ELAN is flexible enough to suit the needs of various types and conventions of annotation, and EAF, the XML data format native to ELAN, is supported as an import and export format by a growing number of tools and frameworks.

ELAN runs well on Windows and Mac OS, while on Linux the situation concerning AV playback needs improvement. Though most of the Java code is platform independent, for media playback, ELAN preferably uses the frameworks that are readily available on a system. ELAN allows to link multiple videos to an annotation document, it can visualize the waveform of audio and the curves of time-series data. In addition it offers several views and editors for annotations.

For documentary linguists, exchanging data with e.g. Toolbox or FLEX is and has been the most common way to create annotations that are both time-aligned and linguistically rich. Although in the meantime work has started on extending ELAN with modules that add semi-automatic interlinear glossing functionality to the program, it is not available yet, and it will not make existing interoperability features superfluous.

ELAN, as a single user desktop application, historically focuses on improving and streamlining local media handling and local data processing. Collaboration is only possible by exchanging files between members of a team. Although some actions can be performed on multiple files (importing and exporting data, searching etc.), which can be seen as operations on a local corpus, most of the work is done on a per document basis.

Only recently the first steps have been made to extend the locally available processing power and algorithms by calling web services. There are two main categories of services that are of interest to ELAN and its users, and the one of interest in this context is a service that works on text, applying parsers and taggers to the input returning new layers of annotations.

2.3 Synergy

The goal of the present project is to integrate annotated data from both ELAN and TC in order to produce a richer set of annotations on the same source material. ELAN will be used for audio and video annotation, while TC will provide interlinear glossing on the language contained in the audiovisual resources. Through the present development, the user will be able to start annotating in either program and then export data and continue annotating in the other program. The data exchange will be done in a standardised XML format improving the already existing TC XML export. It is within the reach of present ELAN development to provide public URIs referring to audio- and video source material. While ELAN's local handling of audio- and video material is state of the art within the present set of technologies, TC is designed for the creation, retrieval and sharing of linguistic data in the form of IGT which is the most common data format in linguistics. The IGT is important since it often is the only structured data available for less-resourced or endangered languages. For these reasons it is important that the digital management of such data finds solid support also in emerging VREs where multi-media annotation, such as facilitated by ELAN will be the other central theme. Designed for the normal linguist, TC can offer simplicity of method to its users together with those features in which online services excel: linking of resources, general openness and a peer-centered design. Complex annotation of media, however, is best handled locally, therefore, a synergy between online and off-line tools is the way we go.

3. A Use-case

The Paunaka Documentation Project (PDP) is funded by ELDP and located at the University of Leipzig. PDP's work comprises the compilation of a corpus of audio and video recordings. The project data is managed with ELAN and archived with ELAR. For the actual data analysis PDP uses both ELAN and TypeCraft where 60% of the data receives a morphological analysis using TypeCraft. Some members of PDP have used ELAN and Toolbox already in earlier projects, while the present project considered three different glossing tools, namely Toolbox, FLEX and TypeCraft. The two main reasons why the project selected TC were that, first of all, TC allows incremental annotation, which is particularly necessary in the beginning of analytic work, and secondly, because TC is an online service and therefore allows the exchange of project data under distributed work.

The project loads all primary data to ELAN where transcriptions and translations are created in one swoop together with the segmentation of the audio. We will come back to primary data processing in ELAN in section 4 since it crucially effects data export. To combine the use of ELAN with the use of an Interlinear Glosser is not new to some project members. In prior projects, Toolbox files had to be exported back into ELAN, not for further morphological processing, but as a prerequisite for archiving (a requirement imposed by some project funding agencies).

Turning now to the combined use of TC and ELAN, the PDP reports that the export of data to TypeCraft is not only tedious but in addition sentence indices and ELAN's time codes 'get lost'. Below we reproduce a project internal description of the present data export procedure used by the PDP:

1. in ELAN select File->Export As->Tab-delimited Text. Select the line to be exported - make sure not to select any other option
2. save
3. open the export in Word
4. use search & replace to delete the line initial
5. if Word recognises the special symbols, copy the text directly to TypeCraft

else:
identify the document for Toolbox, e.g., by \id, open the document in Toolbox, then copy to TypeCraft selecting 'text' as the format

The present data-flow has two major problems: (i) it is too round-about to be effective and (ii) crucial information for the linking of media and text annotation gets lost.

4. Combined system use

In section 2, we have identified where our tools are complementary and where their strong points are, while in section 3, we discussed present problems for data exchange. In spite of these problems, the Paunaka project's decision to combine the use of ELAN and TypeCraft buttresses the usefulness of a working environment that allows to integrate multi-media annotation with linguistic online services. Under the combined use of our tools, TypeCraft will continue to provide the collaborative work environment and access management system as a service online. For our users that means that with a login to TC, the system allows online morphological glossing and collaborative editing, as well as he export of formats not available in ELAN. However, accurate media handling over internet is still problematic, and ELAN, as a local tool, mostly has direct access to the media, which improves the speed and accurateness of the segmentation process. Therefore media handling will in our project remain local. In order to make a combined and flexible use of online service and local tool possible, we need to provide for a seamless data-flow between the online and off-line system. To this end, already implemented changes and planned development will be described in the next section.

4.1 System adaptations

TC was designed as a system to be accessed by humans. Thus, the existing login and export facilities were not directly suited for the exchange of data between TC and ELAN. However, MediaWiki and the existing XML export schema provided a solid basis for developing the initial TC-data-exchange API. By making the web service API available on TypeCraft, we allow users to log in and to retrieve text documents that they have access to, in order to download them and time align them in ELAN. To this purpose the XML schema was extended with several items one of which is the item *Listing of texts without phrases.* This feature will enable TC users to choose transcribed texts (their own or data that they share with a group of other users) for import into their local ELAN where these texts can be aligned with the corresponding media material so that further annotations, important for cross-media comparison and analysis, can be added.
To enable the communication between the two systems we use a simple RESTful (Richardson and Ruby) web service. Access control is provided by enabling the MediaWiki remote API and the exchange of a session id. The following commands regulating export from TC to ELAN have become available to the users of both systems:

- list texts
- export text with 1 to n phrases
- export text with all phrases
- export 1 to n phrases

In ELAN a menu has been added for accessing known web services, such as TC. For the combined system use of TC and ELAN that means that the user can now log in to TC from ELAN. After login, s(he) will be presented with a list of available texts from TypeCraft. Users may have an audio and/or a video already loaded in ELAN or they may download a speech recording directly from TypeCraft. Depending on the text downloaded from TC, ELAN will create one or more tiers to accommodate the imported information. It is crucial for the whole process that the ids of the phrases are retained, otherwise it will be next to impossible to cross-correlate annotations in ELAN with those in TC, effectively turning the import into a one way process.

The described export of already existing TC-data to ELAN will be of particular importance for those TC-users that host speech recordings and corresponding texts on TypeCraft. A combined use of both systems can now provide a better integration of the user's resources and an improved workflow. Text annotations coming from TC can, using ELAN, be confronted with the original recording, and it now becomes possible to correct already existing annotations. Central for connected text is time alignment. Moreover, using ELAN, the user can choose the size of the interval to be time-anchored. Finally, users can enrich already existing TC annotations by adding cross media annotations in ELAN, something that will allow them to identify linguistically important formatives across media.

Turning now to the import of ELAN data to TypeCraft, we have described a laborious version of this exchange in section 3. Paramount for the exchange is that the user of ELAN does not loose information crucial for the alignment of audio/video and text annotation, leading to unwanted separation of recordings and the connected text. In order to facilitate the process, we have updated the TC XML by the following additional items:

(i) *Introduction of the notion Speaker at the phrase level*
So far TC only allowed the storage of speech recordings and transcripts for safe-keeping, yet, their connected processing has not been possible. Neither was it possible in TC to identify speakers or align speaker and utterance. This has now become possible through the present development.

(ii) *Introduction of offset and duration at the phrase level*
This information assures the alignment of text and

audio/video. For the online services we do not plan a visualisation of time-overlap.

In the Paunaka use-case, the project started data processing in ELAN. Focusing only on the here relevant aspects of the project's workflow, this means that intervals were created and first segmentations were made. Project members then create several note tiers, corresponding to the Toolbox \nt. In a first go-through of the material, the researchers use the first note tier to report unknown words or morphemes, or other phenomena that do not allow the initial translation of the material. A common procedure, which we assume is quite general for group work, is to leave comments about initially troublesome features or pointers to interesting grammatical formations so that they at a later point can be selected for integration into the grammar that the project plans to write. Next to the translation and the note tier, it is not unusual to create a loan word tier. A typical working procedure in the early work with ELAN is that transcriptions are corrected together with native speakers, and that notes concerning these corrections are entered on separate tiers, reflecting the comments of different informants. Through the present development we have made sure that the information resulting from the initial processing of data in ELAN is mostly preserved under data exchange. This means that next to the transcribed text and its translation, the speaker identification and the time alignment, also the information coming from note tiers is preserved. However, for the time being, we will have to write all notes, independent of their type, to the already existing note tier in TypeCraft.

We should again mention that the project does not plan the export of audio and video data to TC, that is, the processing of media information will continue to happen locally. Yet, there are ways to make also this data available online for representational purposes, something we will not discuss here further.

In summary, limited reconfiguration using a simple RESTful web service in addition to an improved TC-XML format makes it possible to allow users of our systems to map ELAN tiers (or tier types) to TC tiers and vice versa, to import TC texts into ELAN together with their corresponding speech recordings. Most importantly time alignment information and speaker information is now retained when data is shipped from one system to the other. Manual export and import of the type described by the PDP is through the development presented here no longer necessary.

5. Outlook

Software development profits from an active and articulated user community. Through an approach that targets problems that have become apparent through use-case analysis, small but efficient changes can be made. However, to move towards linguistic VRE, more is needed and mostly these are standards and more standards. With the CLARIN endorsed Metadata Framework (Broeder et.al, 2010) metadata standards have already been addressed. The same is true for a general XML structure for IGT (Bird et.al, 2003). A true challenge is still linguistic modeling. The GOLD ontology, as an RDF resource, is the most suitable starting point for the present project. However, a description of the development of ontology integration, as well as of an improved data parsing and search functionality is well beyond the scope of this paper.

6. Acknowledgements

We would like to thank the Paunaka Documentation Project and especially Lena Terhart for discussion and comments. Paunaka is a South Arawakan language spoken in Bolivia. More information about the Paunaka project can be found on TypeCraft under: Paunaka Documentation Project.

7. References

Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masneri, S., Schneider, D., Tschöpel, S. (2010). ELAN as Flexible Annotation Framework for Sound and Image Processing Detectors. In *Proceedings of the 7th International Conference on Language Resources and Evaluation.*

Beermann, D., Mihaylov, P. (2012). Collaborative Databasing and Resource Sharing for Linguists. In *Proceedings of the 9th Extended Semantic Web Conference, Heraklion, Crete.*

Bird, S., Bow, B., Hughes, B. (2003). Towards a General Model of Interlinear Text. In *Online proceedings of the 3rd E-MELD workshop.*

Broeder, D., Kemps-Snijders, M., Van Uytvanck, D., Windhouwer, M., Withers, P., Wittenburg, P., Zinn, C. (2010). A Data Category Registry- and Component-based Metadata Framework. In *Proceedings of the 7th International Conference on Language Resources and Evaluation.*

Farrar, S., William D. Lewis (2005). The GOLD Community of Practice: An infrastructure for linguistic data on the Web. In *Proceedings of the EMELD 2005, Workshop on Digital Language Documentation: Linguistic Ontologies and Data Categories for Language Resources.*

Richardson, L., Ruby, S. (2007). RESTful Web Services. *O'Reilly Media.*

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: a Professional Framework for Multimodality Research. In *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation.*