Annotation tool Toolbox

How to gloss/annotate in Toolbox

Sebastian Drude

1 Different types of data and annotation

Data. Data is always data *for* or *of* something – usually *a representation of* physical states and events. In linguistics, *primary data* are direct representations or results of a *speech event*, for instance a written text or, in particular, an audio/video recording (a/v.r.) of a speech event.

Annotation. *Annotation of* data is a symbolic representation of properties of the (speech) event represented in the data. (⇒Annotation is data by itself.) In linguistics, the most common and basic types of annotation are a *transcription* (representing formal, in particular sound-related properties) and a *translation* (representing semantic properties) of the linguistic expressions whose utterance is represented in primary data (e.g., an a/v recording).

Global vs. unit-oriented annotation. *Global* or *holistic annotation* represents properties of the event as a whole and is part of the *metadata*. *Unit-oriented annotation* refers to specific parts of the data, in particular, utterances of individual sentences or words or sounds etc.

Secondary and derived data. If unit-oriented annotation is based on primary data (such as a written text or a audio or video recording), it is *secondary data*. Annotation of secondary data would be tertiary data, and so forth recursively – in sum, all u.-o. annotation is *derived data*.

Time-aligned annotation. If unit-oriented annotation is based on data with a timeline, in particular, a media file (a/v.r.), it can be *time-aligned,* i.e., each (piece of) annotation is explicitly associated with the corresponding *chunk* (time-span, segment) of the media file.

Linguistic types of annotations. Annotations differ according to the *types* of properties of the speech event (especially, the uttered linguistic expressions) that are represented – they can be *phonetic, phonological, morphological, syntactic, semantic, pragmatic*, etc., and on each level they can focus on the *units*, or on *structures* of units or *relations* that hold among units.

Coverage of annotation. *Basic annotation* consists of only a transcription and one or several translations and perhaps notes/comments, on a sentence or intonation group level. *Basic glossing* gives additionally information on individual units (usually morphs, sometimes words) such as an individual gloss (indication of meaning or function) and perhaps categorial information such as a part-of-speech tag (or its equivalents on the morphological level). Beyond basic glossing, *advanced glossing* covers one or several of the other levels, from phonetic to pragmatic, and can include, for instance, a phonetic transcription and annotation of the intonation contour, or of the syntactic structure, of grammatical relations, etc.

See “The Leipzig Glossing Rules” for basic glossing (Comrie et al. 2008); and see the proposal “Advanced Glossing” for a systematization of annotation types and their implementation with toolbox (Drude 2003a, 2003b; Lieb&Drude 2001).

2 Annotation tools

Transcriber. Tool for the segmentation and transcription of audio files. – *Pros:* Compatible with MAC, Windows & Linux; very easy to use; produces simple XML-files. – *Cons:* No Unicode input possible; only one line of annotation; no video; no lexicon. (New version not tested.)

ELAN. Tool for the complex annotation of audio and video files. – *Pros:* Compatible with MAC, Windows & Linux; audio and multiple video files; unlimited tiers for different speakers; state-of-the-art; wide user community; XML output (but complex). – *Cons:* Complex tool for beginners (but now: easier transcription mode); no lexicon (yet).

Toolboox. Text-oriented general database tool for linguistic fieldwork with lexicon and texts. – *Pros:* Flexible and powerful; Export to different formats (incl. XML); therefore easy to integrate with other tools; many users. – *Cons:* Too flexible; poor data format "Standard Format"; tricky on MAC; complex to set up; no video and no time-aligning; at end of life-cycle.

FLEX. Extensive linguistic database tool for linguistic fieldwork with lexicon and texts. – *Pros:* Powerful and well-designed; inbuilt ontology and analysis tools; growing user community. – *Cons:* Not flexible (8 tiers); one huge XML database with no good import or export function for texts; Windows only; difficult to configure; no audio, no video, no time-alignment.

Other tools. There are quite some more. For instance: *Praat* is often mentioned, and used for segmenting, but not made for other annotation then phonetic. *CLAN* can annotate audio and video, it is more specific than ELAN, using CHAT or CA (Conversation Analysis) formats, used especially for child language data (CHILDES project). *ANVIL* from the DFKI has a similar scope as ELAN. The *EXMARaLDA Partitur-Editor* from Univ. Hamburg has been widely used for discourse analysis. *Audiamus* and *Eopas* (N. Thieberger) do not create but organize and display annotation data in very useful ways.

	Transcriber	**ELAN**	**Toolbox**	**FLEX**
Complexity	Easy	Complex, but easier modes	Complex to configure	Complex
Audio	Yes	Yes	No (but can play)	No
Video	No	Yes	No	No
Tiers	1 per speaker	Unlimited	Unlimited	Fixed: 8
Lexicon, automatic glossing	No	No (but it is planned)	Yes	Yes
Unicode	No	Yes	Yes	Yes
Data format	Simple XML	Complex XML	Faulty TXT	XML database
Interoperability	Good	Fair	Good	Bad
User community / support	Small?, no support?	Large, good support	Large, fair support	Small, good support
Life cycle	Old (but new version 2011)	Constantly developed	Not officially supported, old	New, being developed

⇒ One needs different tools for (1) time-alignment with audio/video files (Transcriber, ELAN) and for (2) semi-automatic glossing with interoperation with a lexicon (Toolbox, FLEX).

3 Introduction to interlinearizing: Annotation without time-alignment

Annotation of linguistic expressions without any recording (e.g., eliticed data, field notes, written texts: sentences, phrases or even just isolated words) can be done with Toolbox (or Flex). One can use the default setting when starting a new project with **INSTALLTOOLBOXNEWPROJECT###.EXE** (### represents the version number).

The default setting for annotation is in **TEXT.TYP**. It provides a set-up for *basic glossing*: **\REF** Reference (should be unique) – **\TX** Text (sentence) – **\MB** Morphemes (basic form) – **\GE** Gloss (English) – **\PS** Part of Speech (on morphological level) – **\FT** Free translation (English) – **\NT** Notes.

You can type or paste a sentence in the **\TX** field and start "interlinearizing", that is, semi-automatic glossing with parsing and look-up processes defined in the "**INTERLINEAR**" tab in the database type setting. Use **ALT+I** to interlinearize.

Toolbox uses the lexical database to break the sentence up into its morphs and glosses them. If the lexical information is missing (usually marked by asterisks), you have to provide it. Right-click on the unknown item in the **\MB** line, choose "insert" (possibly selecting the right lexical database, if you use more than one; for instance some users prefer a separate database for affixes) and fill in at least the fields needed for interlinearization (in the default setting, the **\PS** and **\GE** fields, which happen to have the same markers as in the text database). Consider using the "alternate (for parse)" field (default: **\A**) and perhaps the "underlying (for parse)" field (default: **\U**), for instance for allomorphs or morpho-phonemic alternations. See the extensive explanations in the Toolbox help and the Shoebox manual (sometimes outdated) for details. (More in the practical training.)

For interlinearization in Toolbox to work properly, the ***language settings*** have to be configured properly. In particular, all letters (also digraphs, such as <***sh***> in English, if you decide to give them a separate place in the alphabet of the language – I recommend not to do so, by the way) have to be correctly represented in the ***sort order***. See the appendix about UNICODE. Always use the same UNICODE representation of a letter if several are possible!

Text-based vs. sentence-based text databases

The default record marker in this setup is **\ID** Text name – the intention is to have only *one* file with *all texts*, each text (with all its sentences) corresponds to *one* record. This setting may not be practical for several reasons, for instance:

- We need separate files for different stories if we want to export them to ELAN.
- If one searches or filters, the hits refer to whole texts.
- If one wants to do advanced glossing, the screen becomes confusing.

It is to be preferred to have each sentence in one record, and each text in one file. For that one should manually remove the **\ID** field in the text file, and must create a new **.TYP** file, copying the original **TEXT.TYP** file (removing all occurrences of "**ID**" etc.). In the new **.TYP** file, do not forget to change the **\+DATABASETYPE *TEXT***, for example to "***SENTENCES***", and to change the database type of the text file (in the first line) accordingly. You can see the changes in the distributed files **SENTENCES.TXT** and **SENTENCES.TYP**, comparing them to **TEXTS.TXT** and **TEXT.TYP**.

4 Annotation with time-alignment: how to prepare and import data

Toolbox is not a tool for *time-linking* (i.e., for defining the time-alignment of each annotation associated with a certain chunk in the mediafile) – it can *play* chunks of audio files, but cannot practically be used to change the *time marks* (start/end times of each chunk). In fact, doing so by hand can lead to serious problems later, especially if chunks overlap.

Therefore, the time-linking has to be done in some other tool, usually together with the first transcription. We focus on Toolbox-relevant features of two tools, ELAN and Transcriber.

ELAN

ELAN can be used for *segmenting* (identification of the time marks of each chunk) and *transcribe* media files. You can export ELAN annotation data to Toolbox format ("Standard Format"), an open it with Toolbox. The results vary depending on the ELAN configuration.

If you use the basic default configuration of ELAN (just one tier "**DEFAULT**", with a transcription), rename that tier to **TX@SPEAKER** (where "**SPEAKER**" is the name of the speaker). Then export the annotation to Toolbox, choosing your **.TYP** file for text databases (in our example above, **SENTENCES.TYP**) as "Shoebox database type" and let ELAN add the tier **REF** (for automatic numbering). The resulting file has the extension **.TBT** (toolbox text) and can directly be opened with Toolbox. Toolbox automatically adds the ELAN-generated fields (e.g., **\ELANBEGIN**, **\ELANEND**, **\ELANPARTICIPANT**) to the settings in the **.TYP** file.

Play chunks of an audio file in Toolbox

In order to play the relevant chunk for a piece of annotation (Shift+F4), Toolbox needs a field in the following form:

```
\wav x:\Toolbox Project\azoamujza.wav 0.742 7.162
```

Any marker can be used (here: **\WAV**). The first element is a file name (with path). The second and third elements are the start and end times, in seconds and with decimal fractions (maximally milliseconds). The elements must be separated by spaces and nothing else.

If you have the time marks as provided by ELAN, you can generate such a field automatically with a *regular expression search and replace* command in an appropriate editor. For instance, with the *jEdit* text editor, you could search for:

```
^(\\ELANBegin +(.+?) *\n\\ELANEnd +(.+?) *)$
```

and replace all matches with:

```
$1\n\\wav X:\\PathToAudioFile\\AudioFileName\.wav $2 $3
```

The details of regular expressions vary quite a bit from system to system, but they are a very powerful means when systematically manipulating text. Look in the help pages of your editor.

You can *hide* (view menu) this generated field and all ELAN generated fields.

Transcriber

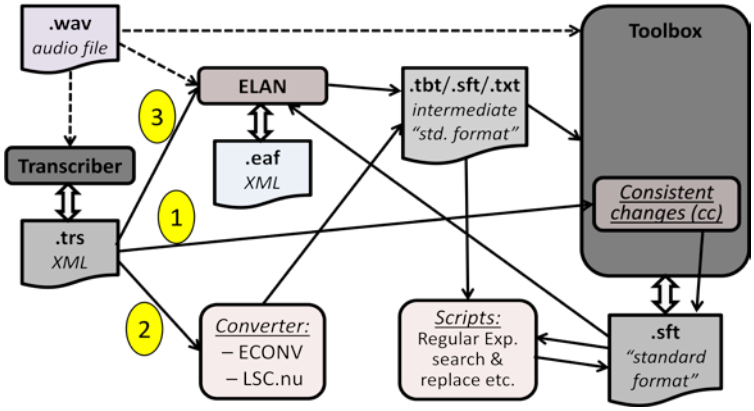
Until recently, the major advantage (ease of use) of Transcriber outweighed its major disadvantage (no UNICODE input). Now, ELAN has the new *transcription mode*, and is a viable

alternative for efficient segmenting and transcribing even for novice users. Still, Transcriber may be an alternative, and has been used by many documentation projects.

In Transcriber, do not forget to set the encoding to UNICODE (UTF-8), although this will not change the strange behaviour not to accept keybord commands for non-Latin1 characters. Use (combinations of) standard Latin1 characters instead. Also, give names to your speakers, and use these; this is particular helpful in the case of overlapping speech.

From Transcriber to Toolbox

There are several ways to import Transcriber generated data into Toolbox, and none is ideal.



In any case, one needs to convert the makeshift (combinations of) characters to UNICODE characters (search and replace), either before or after converting to Standard Format.

(1) The easiest way would seem to be using **Toolbox's transcriber import**, internally applying a *consistent changes* table (from the SIL site http://www.sil.org/computing/toolbox/extras.htm). But there are several problems: The **.trs** file is overwritten without renaming (use a copy!); the **\spkr** and **\sect** are at the wrong position (above **\ref**) ; **\spkr** only appears with a new turn, not for each unit; the speaker names are indicated only once, else "spk1" etc. is given; *overlapping speech* is not preserved. The last point cannot be cured by any additional script.

(2a) There are **dedicated converters**. ***ECONV*** is outdated (and hidden, but still online: http://www.mpi.nl/tg/j2se/jnlp/econv/econv.jnlp). One has to apply both steps, *Transcriber>Toolbox* and *Toolbox>ELAN* (the latter actually produces a Standard Format file with three fields for speaker, start and end points). Still, additional scripts are necessary to adjust field marker names etc.

(2b) The best converter I know is at the site "***Linguistic Software Converters***" by Andrew Margetts: http://linguisticsoftwareconverters.zong.mine.nu. It is completely configurable (choose the names of the markers) and in case of overlapping speech, it generates two entries with identical time marks, which is what can later be processed by ELAN. You can even save your configuration. The result can be directly opened in Toolbox. Careful: the name of the audio file in the field for playing sound is the same as the name of the Transcriber file (which is Transcriber's default but won't work if you rename the audio or transcription file). The only disadvantage is that the converter is a web service only; you have to be on-line to use it.

(3) While offline, the best way is probably to use ELAN as a converter. ELAN can load Transcriber files without problems. Then use the same settings and procedures to export that file to Standard Format. Use scripts to add the audio field and replace possible makeshift letters, and you get a file which is almost the same as if it was generated with Andrew's converter.

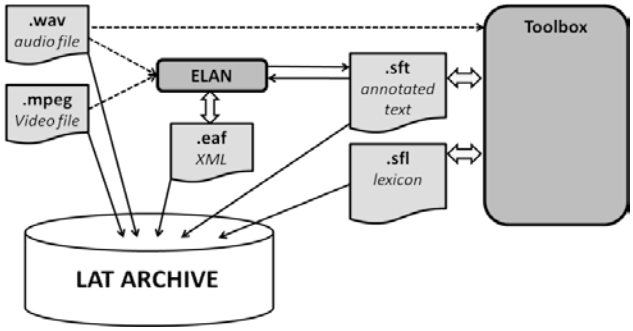
From Toolbox to ELAN and back – "round trip"

After importing a time-linked file (with transcription generated with ELAN or Transcriber) into Toolbox by any of the above ways, use Toolbox to interlinearize it. The same settings as before with non-time-linked annotation should work.

It may be useful to have TWO transcription lines, e.g. one "narrow" transcription, not used for interlinearization, and a normalized one for interlinearization. A script can be used to copy the transcription line in every record. Having a line with normal spacing helps reading.

After interlinarization, that file can be exported to other tools, e.g. to Audiamus or EOPAS, but in particular back to ELAN.

The goal is to have a working "round trip" setting, exchanging files between ELAN and Toolbox.



All annotation files, in particular Toolbox *and* ELAN files should be archived in an ***LAT-archive*** such as the DOBES archive. ELAN files can then be displayed with the *ANNEX* program.

When importing an interlinearized file, ELAN usually preserves the interlinearization correctly, although errors may happen. (Avoid using spaces in glosses or part of speech labels‼ Use dots or underlines instead.) If things should go wrong, ask the ELAN list or the TLA for help.

There is also quite a helpful Toolbox-list: shoeboxtoolbox-field-linguists-toolbox@googlegroups.com

References

Comrie, Bernard; Haspelmath, Martin; Bickel, Balthasar (2008): The Leipzig Glossing Rules. Conventions for interlinear morpheme-by-morpheme glosses. Max Planck Institute for Evolutionary Anthropology & Department of Linguistics of the University of Leipzig. Leipzig. Online at http://www.eva.mpg.de/lingua/pdf/LGR08.02.05.pdf, updated on 2008.

Drude, Sebastian (2003a): Advanced Glossing - a language documentation format and its implementation with Shoebox. In: *Proceedings of the LREC-Conference in May 2002.* Las Palmas: International Workshop on Resources and Tools in Field Linguistics. Paris: ELDA. Online at http://www.emeld.org/workshop/2003/paper-drude.html.

Drude, Sebastian (2003b): Digitizing and Annotating Texts and Field Recordings in the Awetí Project. In: *Proceedings of the third EMELD-Conference on Digitizing and Annotating Text and Field Recordings*. Conference on Digitizing and Annotating Text and Field Recordings. Lansing, Michigan. Lansing, Michigan: E-MELD. Online at http://emeld.org/workshop/2003/paper-drude.html.

Lieb, Hans Heinrich; Drude, Sebastian (2001): Advanced Glossing - A Language Documentation Format. *DOBES Working Papers* 1. Online at http://www.mpi.nl/DOBES/documents/Advanced-Glossing1.pdf.