Annotating Multimodal data of Singing and Speaking

Coralie Vincent

Annotating multimodal data of singing and speaking

Coralie Vincent, Formal Structures of Language Lab, UPL, CNRS-Paris 8 University, France

Introduction

With the advent of complex multimodal corpora, the use of annotation software is now an almost mandatory step in the process of quantitative data analysis. Despite the fact that annotating multimodal data in a widely interoperable and sustainable format is challenging, within the last two decades considerable progress has been made through the development in academia of free and often open-source software allowing annotation of video and audio resources.

The present chapter provides a brief update on the most recent advances made in the field of (free of charge) annotation software, focusing on three multimodal annotation tools: ANVIL (Kipp, 2012), EXMARaLDA (Schmidt & Wörner, 2014) and ELAN (Wittenburg, Brugman, Russel, Klassmann, & Sloetjes, 2006). The second section focuses on the use of ELAN, specifically, how to build a coding scheme, to import (as well as prepare) multimodal data, to annotate, generate statistics on annotations and export the results. The sample research topic presented in this section deals with the way a choir conductor conveys musical indications printed on a score to singers through gesture. The third and final section addresses general considerations regarding interoperability, automation and sustainability of multimodal data annotations.

A brief comparison of three multimodal annotation programs

With the advent of multimodal corpora, many programs have been developed in order to study the temporal relationships of events occurring in different modalities. We focus here on ANVIL (version 6), ELAN (version 5.5.0) and EXMARaLDA[1] (version 1.6) since they are cross-platform[2] (running on Windows, Mac OS X and Linux as they are written in Java, which is wider appeal to range of users), are still being actively developed and widely used in the gesture research community[3]. Our main focus is to present pieces of software versatile enough to allow the analysis of complex data streams such as those studied in the highly interdisciplinary domain of musical gesture analysis (Jensenius, Wanderley, Godøy, & Leman, 2010) in contexts like the AIRS project (Advancing Interdisciplinary Research in Singing).

Previous interesting comparisons of annotation tools for multimodal data include Rohlfing et al. (2006), Duncan, Rohlfing, & Loehr (2013), Perniss (2015) and Glüer (2018), a.o.

Data annotation: why, how?

In recent years, quantitative data seems to be becoming increasingly necessary in music performance studies, in order to "assess expressive aspects of the body movements of the performing musicians" (Goebl, Dixon, & Schubert, 2014) for example. In light of this growing need, annotation proves to be an essential component of analysis in this field.

The annotation workflow process typically includes at least these five steps:

1. Build a coding scheme.
2. Import and synchronise data streams such as audio, video and motion capture.
3. Make annotations or import already existing ones.

4. Query the annotations, generate statistics and save results.
5. Export annotations and excerpts.

User interface and global file organisation

General layout

ANVIL, ELAN and EXMARaLDA are all comprised of:

- at least one media viewer for video and/or audio waveform (plus a 3D motion capture viewer in ANVIL),
- a set of control buttons for media and selection,
- an annotation viewer which contains tiers (annotations cannot overlap within a given tier).

Whereas ANVIL and EXMARaLDA open these viewers in separate windows, ELAN's interface (Figure 16.1) has a single window which contains all panels—this is particularly convenient when working on multiple annotation files at the same time. It seems that older versions of ANVIL included a plugin that allowed viewing a spectrogram but, to the best of our knowledge, this addition is not available anymore in version 6.

The "music-score" interface

Annotation programs are often referred to as "music-score" interfaces (Rohlfing et al., 2006), but the term is only a metaphor. The expression refers to the layout of the interface, where time-aligned annotations unfold horizontally, from left to right, synchronised with various streams (video, audio, motion capture, line graphs of any digital signal), whereas the vertical axis is divided in tiers (or tracks, "staves") containing these annotations. The annotations are arranged as would be notes in an

orchestral score for every instrument.

File organisation

As in non-linear editing systems, media (typically large size video and audio files) and annotations (light size tree-structured XML human-readable text files) are distinct files. Annotation files contain links that point to the media files.

There are subtle differences between ANVIL, ELAN and EXMARaLDA regarding these annotation files. ANVIL separates the annotations (stored in .anvil files) from the way they are organised (the coding scheme, called "specification file", an .asp or .xml file). ELAN stores annotations and the associated coding scheme in the same .eaf file, but allows the coding scheme to be exported as a template .etf for later use with other media or in other studies. EXMARaLDA stores annotations in an .exb file, the annotation specifications in an .xml file and the way files should appear in a "format table" .exf or .xml file. Finally, whereas ELAN allows batch processing (for example, the modification of multiple files), ANVIL and EXMARaLDA work with the notion of corpus, files grouped with the "Project Tool" in ANVIL and Coma (Corpus Manager) in EXMARaLDA.

Common features and differences

Building a coding scheme

The purpose of a coding scheme is to determine what and how to annotate in order to carry out a quantitative analysis and support one's constructs.

This step is absolutely crucial and has to be based on an extensive state of the art. The building of a template and the creation of an annotation manual can be a major section of collaborative research projects (see Allwood, Cerrato, Jokinen, Navarretta, &

Paggio, 2007, for an example of the annotation of feedback, turn management and sequencing phenomena).

In practical terms, the process consists of:

- selecting the categories relevant to the topic of interest (the categories should be exhaustive and independent);
- creating a structure that contains these categories (as tiers) and the relationship between them;
- if necessary, entering a specific limited set of (mutually exclusive) possible values, to limit variability among annotators and annotation files.

The coding scheme can develop further over time (be refined or completed). Making such improvements requires testing and use of the coding scheme.

ANVIL and ELAN allow user-defined hierarchy between tiers, which increases the consistency of coding. The logical relationships between tiers can be defined through "track type" and "group" in ANVIL and "stereotype" in ELAN. Nevertheless, the two programs differ somewhat in this aspect: in ANVIL, a tier can have several attributes (with specific values), whereas in ELAN, these attributes have to be created as tiers (Kipp, 2014). For example, in ANVIL, a user-defined "hand" tier could have a "handedness" attribute with four values: "left", "right", "both-parallel", "both-mirror"). In ELAN, this would translate as a main tier, "hand", parent of a sub-tier, "handedness", for which annotations could be one of four entry values.

EXMARaLDA does not allow this hierarchical level, but offers support for systematic annotation schemes.

ANVIL provides seventeen sample "specifications" under /spec in the installation folder. ELAN sample "templates" are available on the MPI website. EXMARaLDA

provides a template for “annotation specification” files on its website (see Software list below).

Import and synchronise data streams

Audio and video

ELAN is the best of the three programs in its ability to open and read media files, since it integrates several media frameworks, allowing the importation and reading of a wide range of audio and video file formats. EXMARaLDA has recently integrated some of the developments made in ELAN in this area, but it still needs improvement (Schmidt, 2017). Until fairly recently (version 5), ANVIL was behind on this point, and finding an appropriate video codec could be a problem (the codec advised in the help files was the highly compatible but somewhat outdated Cinepak codec, originally written in 1991). This matter has been addressed for video in version 6, that now implements JavaFX Media which supports MPEG-4 and FLV files.
Unlike ELAN, ANVIL cannot open pure audio files, even in version 6. In ELAN, although an audio track can be heard when a video file is played (if there is one), an audio waveform (Figure 16.1 (2)) can only be displayed if it is imported as a .wav file (which can be extracted from the video file or recorded directly on a dedicated high quality sound device).

In general, audio quality is central in studies on vocal production and perception. Therefore, for such studies, only lossless audio (PCM .wav) should be recorded (rather on multitrack), annotated and analysed, avoiding lossy data-compressed files, such as .aac or .mp3 files (as explained in Roads, 2001 and demonstrated in Corbett, 2012).

Motion capture

Motion capture (often referred to as "MoCap") data is "the recording of live motions by tracking a number of key points [markers] in space over time, which are translated into a three dimensional digital representation" (Meredith & Maddock, 2001, p. 1). So far, only ANVIL can natively open and view 3D motion capture data, as well as show trails if a skeleton specification is provided. Kousidis, Pfeiffer, & Schlangen (2013) took advantage of ELAN being open-source to develop a connection between ELAN and InstantReality, Fraunhofer's virtual reality visualisation framework. Unfortunately their work has not been incorporated into the most recent version of ELAN.

Time series

ANVIL and ELAN are able to plot line graphs of time series data, and even to superimpose multiple line graphs in the same plot. This feature allows to open and view practically any kind of data made of time-value points: biometric data such as heart-rate (HRV/ECG), skin conductivity (GSR), facial electromyography (fEMG), pupil dilation, even electroencephalography (EEG); and kinematic data (position, velocity, acceleration and jerk) of body, head, limb (from a MoCap device), but also gaze (from an eye-tracker).

Data synchronisation

When recording multimodal data, synchronisation of the various data streams is essential for quality analysis. Synchronisation can be realised during data gathering, at a cost in terms of equipment, setup complexity and/or software development (see

Papiotis, 2016, p. 56–64 for example), such as clock alignment among audiovisual devices with SPG (Synchronizing Pulse Generator), use of TTL (Transistor-Transistor Logic).

If synchronisation cannot be achieved during the recording session, it can be realised afterwards, if there is no drift among devices over time. ANVIL and ELAN both allow the user to enter offsets for the various media (and time series) involved. Manual synchronisation can be very time consuming, even in the presence of a clap (of a clapperboard, of hands), intended to record an event perceptible by all sensors. But promising developments are made to automate this step (Bannach, Amft, & Lukowicz, 2009).

New and imported annotations

All three programs can import annotations from other software, such as Audacity (an audio multi-track recording and editing program) and Praat (speech analysis and synthesis software, Boersma & Weenink, 2017). In this regard, EXMARaLDA seems to offer the widest range of possibilities, allowing import mostly from other speech transcription tools. ANVIL, ELAN and EXMARaLDA are interoperable to a certain extent: thanks to a joint effort by the respective developing teams, annotations made in ELAN can be directly opened and used in ANVIL and EXMARaLDA and intermediate formats can be used for other combinations of programs (Schmidt et al., 2009).

The main purpose of all three is to create annotations, a very time consuming and detailed work. In all three, annotations are created by defining start times and end times. Then, one can add text content to the selected time span or select possible values from the authorised set. To the best of our knowledge, only ELAN allows on-

the-fly segmentation.

The programs also provide tools to assist in the annotation of specific modalities. For sound (whether music, speech, or bioacoustics), the ability to display a spectrogram can be a very useful and powerful feature. ELAN and EXMARaLDA both allow the user to open the sound of a selection or annotation in Praat, and therefore view a spectrogram of it (in EXMARaLDA, this requires installing the sendpraat subroutine and follow a fairly strict convention for file naming of the audio media, without spaces or special characters). For the head and face, EXMARaLDA provides a useful interface ("Multimodal panel"). So far, only 'Head' (orientation) and 'Eyes' (gaze direction) have been implemented, but work on other facial features ('Brows' and 'Mouth') is still in progress. For motion capture, ANVIL has a 3D viewer and associated motion line graphs with position, velocity and acceleration (which necessitate skeleton specification) can help clarify gesture annotation. Finally, ANVIL and ELAN both allow spatial coding i.e. manually tracking the position of a point of interest in a video (seen as a line graph in ANVIL and as a 2-tuple text in ELAN).

Query of annotations, generation of statistics and saving results

Quantitative analysis of annotations is one of the main aims of the whole process of annotating. ANVIL, ELAN and EXMARaLDA all provide various ways to perform this step, including computing and visualisation of co-occurrences, overlaps and consecutiveness of annotations. All three offer a way to generate simple statistics and various levels of searches (simple to advanced, with the use of constraints and regular expressions, often called "regex") on one or more files, but only ELAN allows a "find and replace" option on one or several files. Whereas ANVIL focuses on data

visualisation (with histograms, transition diagrams and curve comparison) and on the capability to export for statistical software such as SPSS and Statistica (Kipp, 2012), ELAN and EXMARaLDA both allow a more complete analysis of annotations, such as refining a search, saving the search and saving search results.

Only ANVIL and ELAN seem to implement coder agreement and a way to assess probability of certain combinations (with "Association Analysis" in ANVIL and N-grams in ELAN). On the other hand, only EXMARaLDA has a nice interface to construct regular expressions, a catalogue of commonly used search patterns ("Regex Library"), a way to classify search results and linking with metadata. Finally, to further refine data analysis, it is possible to ingest ELAN .eaf annotation files in EXMARaLDA to take advantage of those capabilities not found in ELAN.

Exporting annotations and excerpts

ANVIL, ELAN and EXMARaLDA share the capability of exporting to transcript text and HTML files, as well as saving the current video frame as an image (and ELAN can even save a filmstrip image). But only ANVIL and ELAN offer the possibility to save a video excerpt from a selection.

ANVIL emphasizes its unique capability to export to statistical software (with "Annotation Frame-by-frame") as noted above, and to WEKA, a popular, free, machine learning workbench. ELAN and EXMARaLDA offer most export possibilities to other speech transcription tools and also, interestingly, to the SubRip subtitle format.

Help files and tutorials

When comparing software to decide which one suits the requirements best, it is

important to consider the existence of help files and tutorials, the frequency of updates, completeness of the website, and the size of the user community.

Summary

The elements described above are summarised in Table 16.1.

Table 16.1: Summary of the comparison between ANVIL, ELAN and EXMARaLDA

Software	ANVIL	ELAN	EXMARaLDA
Platforms	Windows / Mac OS / Linux	Windows / Mac OS / Linux	Windows / Mac OS / Linux
Open Source	No	Yes	Yes
Latest version Developed since	Version 6 (August 2017) 2000	Version 5.5 (April 2019) 2001	Version 1.6 (May 2017) 2000
Current Institute (current developer)	Hochschule Augsburg (Michael Kipp)	Max Planck Institute (Han Sloetjes and Olaf Seibert)	Hamburg Centre for Language Corpora (HZSK) since 2011
Main research fields	Multimodality, gesture and sign languages	Gesture and sign languages	Conversation and discourse
Audio and video import Mocap import Time series import	.avi, .flv, .mp4, .mov .asf/.amc, .bvh .csv, .tsv, .txt	.mp3, .wav; .avi, .mov, .mp4, .mpg .x3d (ELAN 4.5 with MINT.tools) .csv, .txt	.wav; .avi, .mov, mp4, .mpg
Annotation helpers	3D viewer and motion curves Video spatial coding (curve)	Synchronous Praat spectrogram Video spatial coding (2-tuple text)	Synchronous Praat spectrogram Multimodal panel (head / eyes)
Analysis and queries (other than simple statistics and searches, from simple to advanced ones, with “regex”, on one or more files)	Data visualisation (histogram, transition diagram and curve comparison) Coder agreement Association Analysis	Search and replace Refine and save a search Save results Coder agreement N-grams	Refine and save a search Save results Regex constructor Regex Library
Exports (other than exporting to transcript text and html, and	Video excerpt Annotation Frame-by-frame for	Video excerpt and filmstrip image To other transcription formats	To other transcription formats To the SubRip subtitle format

saving the current video frame as an image)	statistical software ARFF for WEKA	To the SubRip subtitle format	
Plugins / Interface	Yes / None	Yes / Synchronisation with Praat	No / Synchronisation with Praat
Website Help FAQ Support Mailing list Video tutorials (on Youtube or Vimeo)	http://www.anvil-software.org Not bookmarkable A little scarce[4] On ANVIL's website Email None 6 videos by the author (~40')	http://tla.mpi.nl/tools/tla-tools/elan Ergonomic and bookmarkable Extensive, also on the website None but searchable forum Active forum Yes (To announce new releases) 7 videos by the VIdeo research Lab Aalborg University (~30') 4 videos by Onno Crasborn (~20')	http://exmaralda.org/en Ergonomic and bookmarkable Complete, online .pdf manuals On CLARIN-D portal (German) Email Yes 1 video (~16')

Working with ELAN

The work process in ELAN (detailed in ELAN User Guide Chapter 1, see Software list below) begins with the opening of the media file(s) to annotate (audio or video, up to four video files simultaneously and one .wav file). Once this is done, the main window displays the Media Player(s) (Figure 16.1 (1)), one for each file selected and a Waveform Viewer (Figure 16.1 (2)) if a .wav file has been chosen.

Before going any further, the .eaf annotation file must be saved; it is advisable to continue saving frequently while working (File > Save).

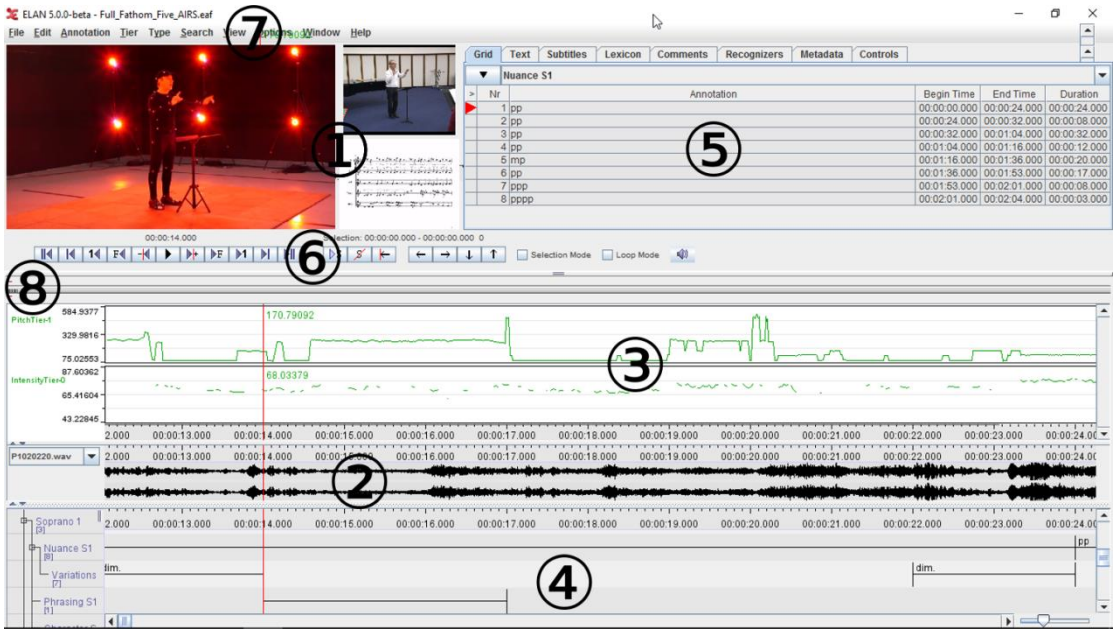
The ELAN interface

The interface language can be modified in Options > Language.

The ELAN window (Figure 16.1) is comprised of the following elements (numbers correspond to elements in the figure):

- Media Player(s) ①
- Waveform Viewer ② (active if a .wav file is opened)
- Timeseries Viewer ③
- Timeline Viewer with tiers containing annotations ④ and a slider on its bottom right to horizontally zoom the time scale in/out
- Alternative, more synthetic, annotation viewers ("Grid", "Text" and "Subtitles") along with other viewers, such as "Comments", "Recognizers"[5] and "Controls" ⑤
- Control buttons for media and selection ⑥
- Menu bar ⑦
- Annotation density viewer ⑧ providing the repartition of annotations along the timeline.

Figure 16.1: The ELAN interface.



Viewing media, waveforms, spectrograms

The video streams to be displayed are selected in View > Media Player and right-clicking on the media player to Detach or Attach them.

If more than one audio file is imported, it is possible to switch between waveforms (View > Waveform). Finally, as already mentioned, Praat can be used within ELAN, to view a spectrogram of a file or selection by right-clicking on the waveform and then clicking Open File in Praat or Open Selection in Praat (unfortunately this is not possible with Sonic Visualiser).

Controlling media players

The Controls tabulation (Figure 16.1 (5)) enables setting global and individual audio levels and global playback speed for all media ("Rate" at the very bottom). The Control buttons (Figure 16.1 (6)) allow navigation in the media (go to begin/end, go

to previous/next scrollview/frame/pixel, play/pause), the annotations (on the right) and the selection (play/clear, go to begin/end).

Working modes

ELAN allows five modes: Options > Annotation Mode, Media Synchronization Mode, Transcription Mode, Segmentation Mode or Interlinearization Mode. These correspond to different steps of the annotation process and ergonomics of the interface.

The "Annotation Mode" is open by default and is the most generally used mode when actually annotating the data. But before annotation, it may be necessary to synchronise media files and time series in the "Media Synchronization Mode". The Transcription Mode and Segmentation Mode are designed to increase the speed and efficiency of transcription and segmentation, respectively. The Interlinearization Mode is specific to parsing annotations and is not likely to be used when working on singing annotation.

Further details on these modes can be found in the ELAN User Guide Chapter 2.5 (see Software list below).

Building a coding scheme

An essential step in the use of ELAN is the construction of a coding scheme, which can be exported as a template .etf for later use with other media or in other studies. The implementation of the coding scheme (or pars of it) is not mandatory before annotation begins.

In the case of choral conducting, we propose the categories and limited set of values given in Table 16.2 (based on Poggi, 2011: 346–347). More general information and

reflection processes about categories to use in music gesture annotation can be found in the literature, e.g. Ashley (2013).

Table 16.2: Coding scheme (tentative) for choral conducting (including modalities, categories, sub-categories and controlled vocabularies, with example)

Modality (grandparent tier)	Category (parent tier)	Sub-category (child tier)	Controlled vocabulary (example)
Music	Bar number	Beats	None
	Tempo	Indication Metronome (from Sonic Visualiser)	None None
	Key		Yes (C min)
	Chords		Yes (G7)
	Soprano Alto Tenor Bass	For each vocal range: Nuances Phrasing and articulation Notes Lyrics	 Yes (*mf*) Yes (>) Yes (D♯) None
Body	Head Trunk Left shoulder Left arm Left hand Right shoulder Right arm Right hand Left leg Left foot Right leg Right foot	For each category: Description Goal Meaning Type	 Yes None None Yes (Enacted Emotion)
Face	Eyebrows Eyes Gaze (if tracked) Mouth	For each category: Description Goal Meaning Type	 Yes None None Yes (Show how to sing)

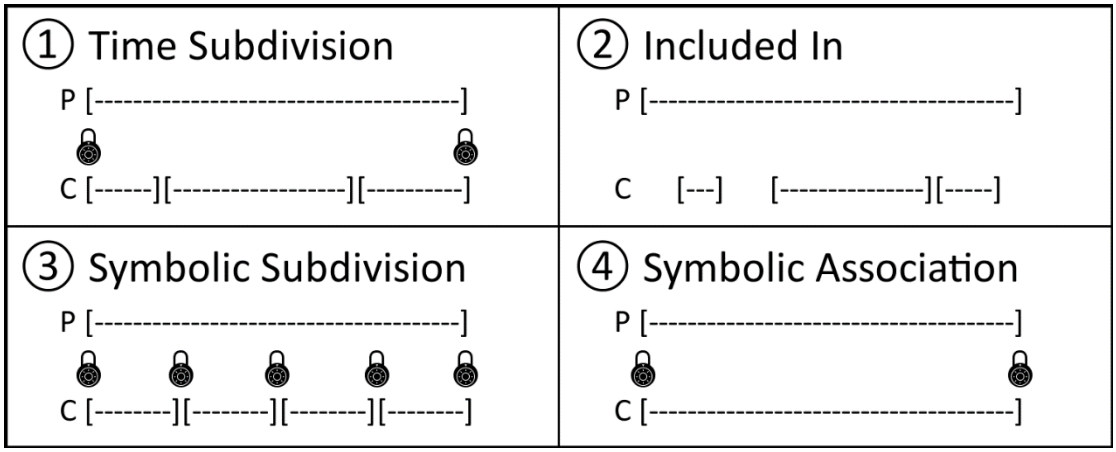
In ELAN, there are two types of tiers:

- independent tiers, whose content is linked to a time interval;
- referring tiers, whose content is linked to a parent tier (itself an independent or referring tier).

The temporal link between a referring tier and its parent tier is constrained by a stereotype. Outside the "None" stereotype for independent tiers, there are four kinds of stereotypes for referring tiers:

- "Time Subdivision" (Figure 16.2 (1)), which allows users to subdivide the annotation of the parent tier into smaller contiguous units;
- "Included In" (Figure 16.2 (2)), which allows users to subdivide the annotation of the parent tier into smaller non-contiguous units;
- "Symbolic Subdivision" (Figure 16.2 (3)), which is similar to "Time subdivision" but subdivides in units of the same duration;
- "Symbolic Association" (Figure 16.2 (4)), which is a one-to-one association without any subdivision.

Figure 16.2: ELAN stereotypes. P: Parent; C: Child.



When a limited set of values, called "Controlled Vocabulary", is defined for a tier, it is linked to the Tier Type. An example of how to create a tier and its Controlled

Vocabulary is detailed in our online tutorial 1[6]. Of course, redundant Controlled Vocabularies (between vocal ranges), such as “Nuance” and “Phrasing and articulation”, are only created once (for example, there is no need to create a “nuance-soprano” Controlled Vocabulary and a “nuance-alto” one).

Saving the coding scheme as a template

Once the coding scheme is entirely implemented in ELAN, it can be saved as an .etf template file for later use with other media or in other studies (File > Save as Template…).

Preparing multimodal data to annotate with ELAN

ELAN can link an unlimited number of audio and video files to an annotation file. It is currently capable of displaying the waveform and simultaneously playing one uncompressed audio file (PCM .wav) and up to four video files, as well as an unlimited number of time series line graphs. Although some file types (such as motion capture or music score) cannot natively be opened in ELAN, there are many possible workarounds to open and display virtually any kind of data, mostly based on the creation of a video file that contains the relevant streams.

Music score video

Scores, if exported as images (from an online repository[7], a scanner, a pdf file or a MIDI file[8]), can be organised as slideshows, which can then be converted to video files. A detailed example is given in our online tutorial 2.1.

Motion capture video

If working with .bvh motion capture files, it is definitely better to annotate with ANVIL, but if the recorded motion capture data is in another format (e.g. .c3d, .trc) and one just wants to have a preview of it in ELAN, Mokka can be used to export the data as an image series. Then a video sequence of the images can be created with FFmpeg or Avidemux. A detailed example is given in our online tutorial 2.2.

Audio

To extract a .wav file (to view its waveform) from a video file, Avidemux is advised. A detailed example is given in our online tutorial 2.3. Be aware that saving a lossy data-compressed audio track from a video file in a .wav uncompressed file will not get the missing data back.

Time series

In ELAN, after a necessary intermediate step for configuration, line graphs can be displayed from .csv files or .txt files containing on each line a time-value(s) pair (values can be in any range) separated with a comma[9]. A detailed example is given in our online tutorial 2.4.

ELAN can also link and directly represent .IntensityTier and .PitchTier files generated by Praat, the software advised for speech analysis. These time series are displayed right after being linked and cannot be configured[10]: ELAN will manage to vertically fit them in the Timeseries viewer.

Music feature line graphs

In the case of singing annotation, it may be better to use Sonic Visualiser[11] (Cannam,

Landone, & Sandler, 2010) rather than Praat[12] to extract music features and export them as text files which will then be linked in ELAN and represented as line graphs. Before using Sonic Visualiser for the first time, and depending on the features to extract, it is recommended to complement it with plugins[13]. Sonic Visualiser is based on the principle of "Layers", where features are shown and their parameters can be set up (one layer per feature).

1) Tempo can be manually extracted by tapping / typing beats (in a "Time Instants Layer" converted to a "Time Values Layer").
2) Pitch is very challenging to compute (Salamon, Gómez, Ellis, & Richard, 2014), but the MELODIA Vamp plugin (Salamon & Gómez, 2012) works fairly well to automatically "estimate the fundamental frequency corresponding to the pitch of the predominant melodic line of a piece of polyphonic (or homophonic or monophonic) music" (MELODIA website).
3) Intensity can be easily created with the "Power Curve: Smoothed Power..." function of the Mazurka Vamp plugin (Earis, 2007).

Then, in Sonic Visualiser, simply select the layer to plot in ELAN and select File > Export Annotation Layer… to export it as a .txt file which contains tab-separated time-value pairs.

Biometrics, kinematics and eye-tracking line graphs

In most cases, biometrics and eye-tracking data (position and pupil size) can be exported as time series and can then be visualised in ELAN. Furthermore, time series can directly be derived up to the third derivative during their importation (as detailed in the next paragraph), which is especially useful for kinematic data: velocity (first derivative), acceleration (second derivative) jerk or jolt (third derivative) can thus be

obtained from positions[14]. Finally, before importing in ELAN, interesting features such as angles between vectors can be computed with a minimum amount of programming[15], in a spreadsheet program like LibreOffice Calc and saved as .csv files.

Importing time series

A detailed example of how to import and view time series in the Timeseries Viewer is given in our online tutorial 2.4.

Importing pre-existing annotations or making new ones

Importing annotations

To import tab-separated time-value pairs as annotations instead of line graphs, there are two steps (plus an optional preliminary step if the data exported is comprised of only two columns, such as tab-separated time-value pairs (as is the case of Sonic Visualiser annotation layers), see a detailed example in our online tutorial 3.1.1). The first step is to transform this .csv file in an ELAN annotation file (see our online tutorial 3.1.2). The second step is to merge the obtained annotation file (from the .csv file) with the actual annotation file (see our online tutorial 3.1.3).

Creating annotations

The steps of the annotation process are detailed in the ELAN User Guide Chapter 4.1 (see Software list below).

Getting coordinates / positions

An annotation can also contain the coordinates of a point of interest in a Media

Player, such as the tip of a pointing finger. The way to get these coordinates is detailed in our online tutorial 3.2. The software Tracker can also be used to get coordinates and import them in ELAN.

Query of annotations, generation of statistics and saving results

Simple statistics

Simple statistics can be accessed and saved in View > Annotation Statistics.

Searching in an annotation file

After annotating a file, annotations can be quantified in terms of number of occurrences and duration (Search > Find (and Replace)…). A detailed example is given in our online tutorial 4.1.

Refining a search, saving it and its results

In the "Search Dialog" window, it is possible to refine a search (Query > Search on current result) and save it as an .eaq file (ELAN annotation query) (Query > Save). Results can be exported as tab-delimited text or as an .eaf file with context by right-clicking on results of a search and choose Export Table As Tab-delimited Text… or Export Results With Context As EAF… respectively.

Exporting annotations and excerpts

ELAN can export annotations in multiple formats (in the main window, File > Export As). Different options are then available, depending on format. For a few formats, it is possible to simultaneously export multiple ELAN files (File > Export Multiple Files As…).

Finally, it is possible to create video excerpts of a selection (File > Export As > Media Clip using Script...), through a command line tool that can extract clips from a video or audio file (such as FFmpeg), using a script file (named "clip-media.txt", located in the ELAN installation folder). It is advised to choose an intra frame as first frame (viewable in Avidemux), otherwise, there may be an offset between the chosen starting point and the actual beginning of the excerpt.

Latest and future developments

ELAN is a very good compromise in terms of ergonomics, ease of use, learning curve and access to help. Its latest developments are oriented towards very useful additional features, such as automation—for segmentation and annotation—and collaborative annotation with the inclusion of a commentary framework (Sloetjes & Seibert, 2016). Future development could include the ability to view motion capture data (as provided in ANVIL), which would be an additional asset to the already great capabilities of the ELAN multimodal annotation software.

Further remarks

Interoperability

"ELAN has found a place in the workflow of a variety of research groups in various disciplines" (Sloetjes, 2011). Its versatility regarding multimodal data (especially import and export functions) and its ease of use (thanks to the very complete documentation) seem to make it the optimal software to occupy the centre of the ecosystem of multimodal analysis of singing. In this way, it allows to take advantage of the "best of every world" (switching from one program to another, e.g. to view

MoCap data in ANVIL after having annotated the associated video in ELAN).

Automation (unavoidable with increasingly large datasets)

Apart from annotating, it should be emphasised how important are the previous steps in multimodal data analysis, namely recording and quality assessment of the data to automate analysis, segmentation and transcription[16]. For example, it is highly advised to do a multitrack recording of the various audio sources (with multiple and as separated as possible microphones) for a successful automatic feature extraction. For researchers who are not familiar with data gathering, it is a good idea to take into consideration beforehand advice on recording audio (Vincent, 2012) and video (Perniss, 2015). As automated detection libraries (such as OpenPose and OpenFace) are becoming more accessible and functional, it is essential to be aware that these libraries are complementary to naked-eye observation and quantitative methods. But their results should always be questioned and scrutinised[17], especially if quality standards of the input data are not met.

Sustainability

Finally, at a time when large data sets and open science are becoming more and more common, sustainability (in the sense of what and how the data is described and analysed) is crucial[18]. The issues at stake are the possibility to open annotations and data long into the future, the possibility to make the analytic procedure "transparent and available to critique" (Duncan et al., 2013), and the reproducibility of analyses (a point already raised in the field of automatic music transcription by Benetos, Dixon, Giannoulis, Kirchhoff, & Klapuri in 2013).

Conclusion

We presented in this chapter a comparison of ANVIL, ELAN and EXMARaLDA, three programs that have now reached maturity. We then emphasised the use of ELAN, with a focus on its capability to interoperate with a set of software oriented towards the multimodal analysis of singing. While our own work is motivated by the study of choir conductors conveying emotion, the concepts and tools introduced are naturally transferrable to many themes; the use of ELAN in our specific context situation can be easily extrapolated to multiple (types of) researches involving singing in its multimodal aspects.

Much ground has been covered since the publication of the seminal paper by Bigbee, Loehr, & Harper (2001) on multimodal annotation. But new challenges arise with the use of machine learning algorithms to segment and classify. Even though these automated techniques highly reduce the amount of time necessary to annotate, for the time being it is still necessary for researchers to complement these tools with “naked eye” and “naked ear” observations (Jensenius et al., 2010).

Endnotes (see end of file)

Acknowledgments

The author would like to thank Dominique Boutet (project leader), Marion Blondel, Fanny Catteau, all the other members of the CIGALE project (funded by LabEx ARTS-H2H), and Ariel Alonso (choir conductor) for their precious collaboration, as well as the two reviewers and Dana Cohen for their extremely valuable comments.

References

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation*, *41*(3–4), 273–287. doi:10.1007/s10579-007-9061-5

Ashley, R. (2013). 106. Bodily interaction (of interpreters) in music performance. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Body - Language - Communication* (Vol. 38.2, pp. 1432–1440). Berlin: De Gruyter Mouton. doi:10.1515/9783110302028

Bannach, D., Amft, O., & Lukowicz, P. (2009). Automatic event-based synchronization of multimodal data streams from wearable and ambient sensors. In P. Barnaghi, K. Moessner, M. Presser, & S. Meissner (Eds.), *Lecture notes in computer science: Smart sensing and context* (Vol. 5741, pp. 135–148). Berlin, Heidelberg, NewYork: Springer. doi:10.1007/978-3-642-04471-7_11

Benetos, E., Dixon, S., Giannoulis, D., Kirchhoff, H., & Klapuri, A. (2013). Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, *41*(3), 407–434. doi:10.1007/s10844-013-0258-3

Bigbee, T., Loehr, D., & Harper, L. (2001). Emerging requirements for multi-modal annotation and analysis tools. *Proceedings of Eurospeech 2001 - Scandinavia* (pp. 1533–1536).

Boersma, P., & Weenink, D. (2017). Praat: Doing phonetics by computer [Computer program]. Version 6.0.29, retrieved 6th June 2017 from http://www.praat.org/

Cannam, C., Landone, C., & Sandler, M. (2010). Sonic Visualiser: An open source

application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 1467–1468). New York, NY, USA: ACM. doi:10.1145/1873951.1874248

Corbett, I. (2012). What data compression does to your music. Retrieved 6 June 2017, from http://www.soundonsound.com/techniques/what-data-compression-does-your-music

Duncan, S., Rohlfing, K., & Loehr, D. (2013). 66. Multimodal annotation tools. In C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & S. Tessendorf (Eds.), *Body - Language - Communication* (Vol. 38.1, pp. 1015–1022). Berlin: De Gruyter Mouton. doi:10.1515/9783110261318.1015

Earis, A. (2007). An algorithm to extract expressive timing and dynamics from piano recordings. *Musicae Scientiae, XI*(2), 155–182. doi:10.1177/102986490701100202

Glüer, M. (2018). 13 - Software for coding and analyzing interaction processes. In E. Brauner, M. Boos, & M. Kolbe (Eds.), *The Cambridge handbook of group interaction analysis* (pp. 245–274). Cambridge, UK: Cambridge University Press. doi:10.1017/9781316286302.014

Goebl, W., Dixon, S., & Schubert, E. (2014). Quantitative methods: Motion analysis, audio analysis, and continuous response techniques. In D. Fabian, R. Timmers, & E. Schubert (Eds.), *Expressiveness in music performance – Empirical approaches across styles and cultures* (pp. 221–239). Oxford, UK: Oxford University Press. doi:10.1093/acprof:oso/9780199659647.003.0013

Humphrey, J. D., & Delange, S. L. (2004). *An introduction to biomechanics: Solids and fluids, analysis and design*. New York, NY, USA: Springer-Verlag. doi:10.1007/978-1-4899-0325-9

Jensenius, A. R., Wanderley, M. M., Godøy, R. I., & Leman, M. (2010). Musical gestures: Concepts and methods in research. In R. I. Godøy & M. Leman (Eds.), *Musical gestures: Sound, movement, and meaning* (pp. 12–35). New York, NY, USA: Routledge.

King, E., & Ginsborg, J. (2011). Gestures and glances: Interactions in ensemble rehearsal. In E. King & A. Gritten (Eds.), *New perspectives on music and gesture*. doi:10.4324/9781315598048-17

Kipp, M. (2012). Chapter 21 - Multimedia annotation, querying, and analysis in ANVIL. In M. T. Maybury (Ed.), *Multimedia information extraction: Advances in video, audio, and imagery analysis for search, data mining, surveillance and authoring* (pp. 351–368). John Wiley & Sons. doi:10.1002/9781118219546.ch21

Kipp, M. (2014). Chapter 21: ANVIL: The video annotation research tool. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 420–436). Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/9780199571932.013.024

Kousidis, S., Pfeiffer, T., & Schlangen, D. (2013). MINT.tools: Tools and adaptors supporting acquisition, annotation and analysis of multimodal corpora. In *Proceedings of Interspeech 2013* (pp. 2649–2653). Lyon, France: ISCA.

Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17–29).

Meredith, M., & Maddock, S. (2001). *Motion capture file formats explained* (No. CS-01-11). Sheffield, UK: Department of Computer Science, University of Sheffield.

Papiotis, P. (2016). *A computational approach to studying interdependence in string*

quartet performance. Unpublished PhD dissertation, Universitat Pompeu Fabra, Barcelona, Spain.

Perniss, P. (2015). Collecting and analyzing sign language data: Video requirements and use of annotation software. In E. Orfanidou, B. Woll, & G. Morgan (Eds.), *Research methods in sign language studies: A practical guide* (pp. 55–74). .Blackwell-Hoboken, NJ: Wiley

Poggi, I. (2011). Chapter 25. Music and leadership: The choir conductor's multimodal communication. In G. Stam & M. Ishino (Eds.), *Integrating gestures: The interdisciplinary nature of gesture* (pp. 341–354). Amsterdam: John Benjamins.

Roads, C. (2001). *Microsound*. Cambridge, MA, USA: The MIT Press.

Rohlfing, K., Loehr, D., Duncan, S., Brown, A., Franklin, A., Kimbarra, I., … Wellinghoff, S. (2006). Comparison of multimodal annotation tools. *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion (Discourse and Conversation Analysis)*, *7*, 99–123.

Salamon, J., & Gómez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech and Language Processing*, *20*(6), 1759–1770. doi:10.1109/TASL.2012.2188515

Salamon, J., Gómez, E., Ellis, D. P. W., & Richard, G. (2014). Melody extraction from polyphonic music signals: Approaches, applications, and challenges. *IEEE Signal Processing Magazine*, *31*(2), 118–134. doi:10.1109/MSP.2013.2271648

Schmidt, T. (2017). New official version – exmaralda.org. Retrieved 12 May 2019, from https://exmaralda.org/en/2017/04/27/new-official-version/

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J., Kipp, M., Loehr, D., … Sloetjes, H.

(2009). An exchange format for multimodal annotations. In M. Kipp, J.-C. Martin, P. Paggio, & D. Heylen (Eds.), *Multimodal corpora* (pp. 207–221). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-04793-0_13

Schmidt, T., & Wörner, K. (2014). Chapter 20: EXMARaLDA. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 402–419). Oxford, UK: Oxford University Press. doi:10.1093/oxfordhb/9780199571932.013.030

Sloetjes, H., Wittenburg, P., & Somasundaram, A. (2011). ELAN - Aspects of interoperability and functionality. *Proceedings of Interspeech 2011* (pp. 3249–3252). Florence, Italy.

Sloetjes, H., & Seibert, O. (2016). Measuring by marking; the multimedia annotation tool ELAN. In A. Spink, G. Riedel, L. Zhou, L. Teekens, R. Albatal, & C. Gurrin (Eds.), *Proceedings of Measuring Behavior 2016* (pp. 492–495). Dublin, Ireland.

Vincent, C. (2012). *Audio recording dos and don'ts*. Presented at the AIRS 4th Annual Meeting, Charlottetown, Canada. Retrieved from https://hal.archives-ouvertes.fr/hal-02126747

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., & Sloetjes, H. (2006). ELAN: a professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)* (pp. 1556–1559).

Software and plugins (free and mostly open-source)

Reviewed software:

- ANVIL: video and motion capture data annotation. https://www.anvil-

software.org/

- ELAN: video and audio annotation. https://tla.mpi.nl/tools/tla-tools/elan/
    - Sample “templates”: https://tla.mpi.nl/tools/tla-tools/elan/thirdparty/
    - ELAN User Guide Chapter 1: https://www.mpi.nl/corpus/html/elan_ug/ch01.html
    - ELAN User Guide Chapter 2.5: https://www.mpi.nl/corpus/html/elan_ug/ch02s05.html
    - ELAN User Guide Chapter 4.1: https://www.mpi.nl/corpus/html/elan_ug/ch04.html
- EXMARaLDA: video and audio annotation. https://exmaralda.org/en/
    - Template for annotation specification files: http://exmaralda.org/en/utilities/

Other software:

- Audacity: audio multi-track recording and editing. https://www.audacityteam.org/
- Avidemux: basic video editing (with batch processing capabilities). http://avidemux.sourceforge.net/
- FFmpeg: video and audio converting (in command line). https://ffmpeg.org/
- LibreOffice Calc: spreadsheet. https://www.libreoffice.org/discover/calc/
- Mazurka (Vamp plugins): spectral visualisation and feature extraction (from the Mazurka project). http://www.mazurka.org.uk/software/sv/plugin/
- MINT.tools (plugin compatible with ELAN 4.5): connection between ELAN and InstantReality, Fraunhofer’s virtual reality visualisation framework. http://www.dsg-bielefeld.de/mint/
- MELODIA (Vamp plugin): fundamental frequency estimation of the

predominant melodic line of a piece of music. https://www.upf.edu/web/mtg/melodia

- Mokka: motion kinetic and kinematic viewing and analysis. https://biomechanical-toolkit.github.io/mokka/
- MuseScore (used in the online tutorial): music notation. https://musescore.org/
- OpenFace: face recognition library. http://cmusatyalab.github.io/openface/
- OpenPose: real-time multi-person keypoint detection library. https://github.com/CMU-Perceptual-Computing-Lab/openpose
- Praat: speech viewing, analysis and synthesis. http://www.fon.hum.uva.nl/praat/
- Sonic Visualiser: music viewing and analysis. https://www.sonicvisualiser.org/
- Tracker: video analysis and modelling tool. https://physlets.org/tracker/
- WEKA: machine learning workbench. https://www.cs.waikato.ac.nz/ml/weka/

Glossary

- Annotation: the marking of a segment with a starting point and an end point (typically, with a temporal interpretation), and one or more text labels (with no fixed interpretation). (based on Schmidt et al., 2009)
- Coding scheme (annotation scheme): an explanatory system supplying information about the annotation practices followed, and the explicit interpretation for the annotation. (based on Leech, 2005)

- Intra frame (key frame): a frame which contains an entire image in a compressed video stream, contrary to a Predicted frame.
- Lossy data compression: a process of irrevocably reducing the amount of data required to represent digital content.
- Plugin: a separately installed module that adds specific features to an application.
- Skeleton (3D skeleton): a hierarchical set of interconnected rigid bones whose positions are estimated from motion capture skin markers or body poses.
- Spectrogram: a visual representation of the frequency contents of a signal over time showing the strength of the spectral density expressed as colours or greyscale values.
- Waveform (oscillogram): a visual representation of a signal amplitude over time.

---

[1] More precisely, as EXMARaLDA is a suite of programs, we are focusing on Partitur Editor (for annotating) and EXAKT (for querying and analysing).

[2] Unless specified otherwise, all the software mentioned in this chapter are cross-platform and open-source.

[3] Some proprietary programs such as The Observer XT (by Noldus) or INTERACT (by Mangold) also exist and are used in the music gestures community (see for example King & Ginsborg, 2011) but are not detailed in this chapter.

[4] ANVIL help can be scarce about the software itself but it is very detailed about processes that precede or are linked to the use of ANVIL (converting videos, importing time series—plot data—, exporting from Praat to import in ANVIL, theoretical background about measuring coding agreement...).

[5] These tabs can be removed: uncheck them in View > Viewer.

[6] The examples indicated with the icon  are detailed in an online tutorial available at: https://hal.archives-ouvertes.fr/hal-01964707

[7] https://www.loc.gov/notated-music/ (images); http://gallica.bnf.fr/html/und/partitions/partitions (images); http://www.bac-lac.gc.ca/eng/discover/films-videos-sound-recordings/sheet-music-collection/Pages/sheet-music-collection.aspx (pdf files)

[8] https://musescore.com/sheetmusic?text=Search+for+sheet+music

[9] .txt files containing more than two columns separated by a tabulation (not a space) can also be

displayed after configuration.

[10] .tsv or .txt files with only two columns (one for time, one for value) and a tabulation or a space as separators are also displayed right after being linked and cannot be configured (error msg “There haven't been any configurable time series sources defined”).

[11] Tutorials are available online: http://www.sonicvisualiser.org/videos.html; http://www.charm.rhul.ac.uk/analysing/p9_1.html

[12] But interested musicologists can still have a look at this tutorial to use Praat: https://wimvandermeer.wordpress.com/praat-manual-for-musicologists/

[13] http://www.vamp-plugins.org/download.html is a good place to start looking for Vamp plugins.

[14] See Humphrey & Delange, 2004, p. 335–336.

[15] *Id.* p. 368–371.

[16] More generally speaking, one should always think (at least) one step further (if not consider the whole dataflow) when gathering and analysing data.

[17] As highlighted by Duncan et al. (2013): “[…] tools such as those used for analysis of multimodal communication are not ‘neutral’ […]”. One could think that this problem would be solved by automation. But nothing is less certain, as algorithms may induce a bias, depending on the model they are using, the way they are implemented and so on.

[18] This is one of the reasons why the writing of an annotation manual is so strongly advised.