Gesprächsforschung - Online-Zeitschrift zur verbalen Interaktion (ISSN 1617-1837)
Ausgabe 7 (2006), Seite 99-123 (www.gespraechsforschung-ozs.de)

Comparison of multimodal annotation tools — workshop report

Katharina Rohlfing / Daniel Loehr / Susan Duncan / Amanda Brown / Amy Franklin / Irene Kimbara / Jan-Torsten Milde / Fey Parrill / Travis Rose / Thomas Schmidt / Han Sloetjes /Alexandra Thies / Sandra Wellinghoff

Abstract
In recent years, dozens of tools have become available for annotation of digital audio-video data. For a researcher looking for an annotation tool, it is difficult to decide about its usefulness and usability. In this paper, essential information about some of the available tools is summed up as a result of a workshop at which developers, user experts, and researchers interested in using these tools met. At this forum, these tools' strengths and weaknesses for specific annotation and analysis purposes were discussed.

Keywords: multimodal annotation, software, usability

1 Motivation and format of the workshop

In recent years, dozens of tools have become available for annotation of digital audio-video data. At first glance, many of them look promising and offer a variety of useful features. Yet for the gesture researcher hoping to use such a tool, it can be difficult to determine whether a particular one is suitable for her or his data set, research question, or available computer. To decide about usefulness and usability, it is necessary to know about the ease of use, strengths/weaknesses for specific annotation purposes, and the type of data or analysis the tool is designed for — knowledge that is usually gained only after becoming an expert in the use of a particular tool. The goal of the workshop at the Second Congress of the International Society for Gesture Studies in Lyon (June 15-18, 2005) was, thus, to present information about and demonstrations of some of these tools and to offer a forum for developers, user experts, and researchers interested in using these tools.

The workshop lasted three hours. For the first two hours, experienced users of half a dozen tools presented their expertise using the tool in two pre-workshop exercises:

1. ***Freestyle assignment,*** to reveal the strengths of each tool (i.e., what it is designed for), the users had annotated and analyzed a data set of their own choice, within their own preferred research topic.
2. ***Compulsory assignment,*** to reveal possible weaknesses of each tool, the users also annotated a common data set on a common research topic provided before the workshop. The data set consisted of audio-video clips of the same subject in four elicitations (see also Figure 1):
   a) free conversation (4-person)
   b) storytelling (2-person)
   c) collaborative planning (2-person)
   d) route description (2-person)

Figure 1: Four elicitations in the compulsory assignment: free conversation, storytelling, collaborative planning, and route description.

Users examined how the target subject's speech and gesture differed across these elicitations. Their analysis provided the basis for a global comparison across the different tools. The idea behind this exercise was that many tools are ill-suited for purposes for which they were not designed. For example, one tool may not handle the common video data (e.g., too long or wrong format); another may not support a certain type of analysis (e.g., one requiring fine-grained assessment of gesture-speech synchrony or annotation of multi-party interactions). Information about limitations such as these is not generally advertised by tool developers but would be of value to potential tool users.

In addition to the individual tool reports, the workshop organizers (Daniel Loehr, Susan Duncan, and Katharina Rohlfing) developed an overall tool comparison, which can be viewed in table format in Section 4. Results of this workshop are captured in an on-going web forum, a resource for potential tool users to consult in the future:
http://vislab.cs.vt.edu/~gesture/multimodal/workshop/index.php

The final hour of the workshop was a hands-on session, where participants got the chance to try out the tools with both the experienced users and tool developers on hand. The list of tools we were able to compare is:

1. Media and text editors
2. ANVIL Version 4.5 (http://www.dfki.de/~kipp/anvil/)
3. ELAN Version 2.4.1 (http://www.mpi.nl/tools/elan.html)
4. EXMARaLDA Version 1.3.2. (http://www.rrz.uni-hamburg.de/exmaralda/index-en.html)
5. TASX (http://medien.informatik.fh-fulda.de/tasxforce/TASX-annotator)
6. MacVisSTA (http://vislab.cs.vt.edu/~rtr/)

2 Tool evaluation

In Section 3, experienced users and developers give a brief description about a particular tool. Since the table presented in Section 4 gives an extended overview according to a variety of categories, the users and developers were asked to focus on the main idea/purpose of the tool, their personal opinion, and the usability of the tool. The following usability criteria were suggested:

a) How difficult is it, and how long does it take to learn to use it?

b) How quickly can the data be annotated? (speed of execution)

c) Mental load: Does a user have to think carefully and have much of information in mind while using this tool?

d) How often do errors occur, and how serious are they? This question accounts for both sides of the interface: the errors in the form of break-downs of the system or its limitations as well as popular "errors" generated by users when they annotate data.

3 Presentation of the tools

3.1 Media and text editors *(Susan Duncan)*

Among tools that support analysis of multimodal discourse data, media and text editors, though regarded as "low-tech", may still have the broadest user base, though they are generally regarded as "low-tech". This is true despite the increasing availability of computer programs that integrate visualization, annotation, and analysis capabilities for digitized data of many types (audio-video, motion-tracking, biometric) in interactive coding interfaces. As further tool descriptions in this report demonstrate, many of these software interfaces are highly user-configurable and offer analytic capabilities well beyond those afforded by media and text editors.

3.1.1 Main idea/purpose of the tools

By text editors, we mean word processing software and also spreadsheet calculators. Many researchers use the latter to transcribe intervals of speech into records; the other fields are for identifiers and various labels categorizing the interval of speech and any co-occurring gestures in some way relevant to an analytic goal. In word processing documents, annotations are made to transcribed speech. These signify the co-occurrence of other behaviors with the speech. For example, square bracketing on intervals of transcribed speech on the left side of Figure 2 is an annotation convention that indicates a gesture of some sort co-occurs with this speech. By media editors, we mean either professional-grade VCR editing decks such as the Sony EVO-9650 pictured in the upper right of Figure 2 or audio-video editing software such as Adobe Premiere Pro™ or Apple Final Cut™, a screen shot of which appears in the bottom right. Both platforms, the one Hi8 tape-based, the other digital media-based, provide data handling functions that many researchers who work on multimodal discourse data regard as essential. Particularly, all permit the playback of audio-video data at varying slow motion speeds with clear video images and access to the concurrent audio track, even at frame-by-frame speed. Consumer-grade VCRs do not provide this latter playback capability, nor does software such as QuickTimePro™. Rotational jog/shuttle controllers are available for both platforms. The one pictured at the middle right of Figure 2 is a Contour Design ShuttlePRO™. These facilitate moving forward and backward at the various speeds needed for fine-grained observation of co-occurring verbal and nonverbal behaviors.

3.1.2 Usability

Some of the most widely cited research on multimodal discourse continues to be carried out largely using these multi-purpose technologies. There are several reasons for this: (1) ease of use, (2) learnability, (3) reliability, (4) ready technical support, (5) corpus accumulation uninterrupted by software obsolescence, and (6) long-term and wide access to legacy corpora. Each of these is to some extent a consequence of the fact that media and text editors are backed by commercial concerns because of the market potential of these technologies. This contrasts with the situation of many of the specialized visualization and annotation tools, often developed and maintained by individual researchers specifically for multimodal discourse research. To the above six usability features of media and text editors, we add a non-obvious advantage of simple text editors over other interfaces that support the accumulation of observations. This is that an annotated text transcript of a discourse provides an integrative visualization of raw and coded metadata covering a large extent of connected discourse, one that is intuitive and readily cross-comparable with the transcripts of other discourses.

Figure 2: Left: Portions of speech transcripts partially annotated for nonverbal behavior. Right: Options for working with tape versus digital media.

The sort of analysis described here is one that is concerned less with aggregating occurrences of particular *a priori* categorized behaviors for the purpose of summing across entire discourses; it is concerned more with detecting, for example, large scale patterns that unfold across discourses or small scale sequential dependencies whose existence the researcher may have no reason to expect, *a priori*. As of this writing, we know of no software interfaces capable of generating integrative visualizations of the sort we mean here, those in response to user queries of accumulated codings in a database (although see Schmidt's report on EXMARaLDA, below).

3.1.3 Drawbacks

Compared to "interactive music-score"-type interfaces such as Anvil, Elan, and TASX, media and text editors impose two serious limitations on observation and analysis of multimodal discourse. The first is illustrated in the partially annotated interval of free conversation, running down the left side of Figure 2. During several turns by participants B, C, and D, participant A (in red) is continuously shaking her head; a long interval during which she performs a single nonverbal behavior. Yet the text document format requires division of this behavior across several lines of the transcript, giving the appearance of an iterative rather than continuous behavior. When discourse data consist of multiple interacting participants, speaking and producing various nonverbal behaviors for short and long intervals and overlapping with one another, the annotated text document format constrains observation and analysis. The following "music-score" interfaces show real advantages for discourse data of this complexity. The second serious limitation concerns data aggregation to support quantitative analyses. Search and query of a database consisting only of annotated speech transcripts must be done "by eye", augmented with the minimal search capabilities provided by word processing software. This can be less of a concern when spreadsheet calculators are used for transcribing and accumulating observations, provided the analyst gives adequate forethought to the field structure and content of records in the database.

3.2 Anvil: The video annotation research tool

(Daniel Loehr and Amy Franklin)

Anvil (Kipp 2001, 2004) allows for flexible, intuitive annotation at the expense of a moderate learning curve. Figure 3 shows a screen shot as used in the Lyon workshop.

In Figure 3, the top middle window shows the video, while the large window at the bottom, the "annotation board", contains user-defined, time-based annotations in the typical "musical score" layout. The horizontal axis is time (in video frames), and the vertical axis is a collection of user-defined "tracks", each for a phenomenon of interest. The annotation board and video are time-aligned such that moving the red vertical line (the "playback line") in the annotation board advances or rewinds the video and vice versa. The user creates annotations by clicking a start-point in the desired track at the desired time, advancing or rewinding the video as quickly or slowly as desired (even frame-by-frame), and clicking again in the track (or using a keyboard shortcut) to mark the end-point of the annotation’s interval. Further information about the annotation can then be entered by setting user-defined categories (often with pulldown menus or radio buttons) or entering free text. The topright window in Figure 3 displays such information about a selected annotation. Finally, the topleft window displays program execution status as well as video playback controls.

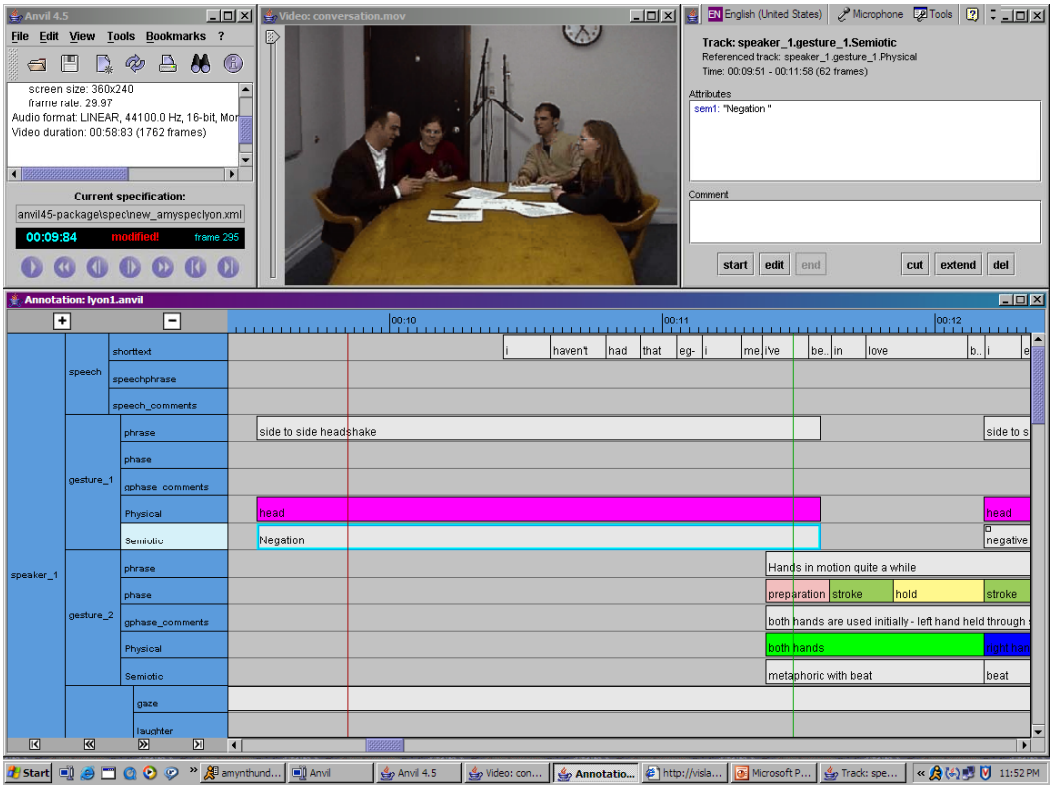

Figure 3: A screen shot of Anvil working with the Lyon workshop "compulsory" data.

Technical details, capabilities, and limitations of Anvil can be found in the table (Section 4) accompanying this paper. This discussion will focus on high-level pros, cons, and usability issues:

3.2.1 Usability

A drawback of Anvil is the certain amount of technical savvy required to install the software (including the underlying Java and Java Media Framework which it relies on), to ensure that the user's machine has the correct video codec installed, and to obtain a video actually loaded into Anvil. In fairness, much of this is outside the developer's control and is common to all such tools. The software does crash occasionally, but technical support is very responsive, and the tool has been steadily improving in stability. Another issue is that the user's preferred annotation types must be specified in XML. XML is not difficult to learn, but annotation schemes of any real interest will be moderately complex and defining them can be tricky. Most users will adopt an existing XML definition file (several are provided with the software) and then make modifications.

Once over the learning curve, however, Anvil provides great flexibility in defining annotations and an intuitive, graphical interface for quickly making annotations. The annotations can be hierarchically grouped for logic's sake and visually minimized or collapsed for visualization's sake. Annotations can be added, deleted, or re-defined at any time (again, by working with the sometimes difficult XML files), and the data can be re-loaded without having to start the annotation process all over again. A great feature is that annotations (including

waveforms and pitch tracks) can be imported from popular speech annotation software like Praat and XWaves, and annotations can be exported as time-stamped files for statistical processing in packages like Excel, SPSS, Theme™, or user-supplied scripts. Even without statistical processing, Anvil's visual interface allows the user to visually scan the annotations for patterns. There is also a built-in search feature to find phenomena of interest, which can then be bookmarked for rapid retrieval.

In sum, Anvil is a solid choice for the multimodal researcher willing to invest some time installing and learning the tool. It has a wide user base and has been successfully used for a number of projects, as described on the Anvil web site (see Section 4).

3.3 ELAN *(Amanda Brown and Han Sloetjes)*

3.3.1 Main idea/purpose of the tool

Elan is a linguistic annotation tool for the creation of text annotations for audio and video files. The annotations can be grouped on multiple layers or tiers that are part of tier hierarchies. The annotation values are Unicode characters, and the annotation document is saved in XML format. Available for Windows, Mac, and Linux users, Elan has been designed for speech and gesture research and is increasingly used in sign language studies. Its main advantages are that it is free; it has a relatively shallow learning curve; its interface is user-friendly, and it is constantly being improved in response to user suggestions.

3.3.2 Usability

Elan can be freely downloaded by PC, Mac, or Linux users from the Max Planck Institute website (www.mpi.nl/tools/elan.html). Once downloaded, users can proceed in one of two directions. If completely new to annotation tools for audio-visual data or with only basic knowledge/experience, users can begin with the new "Getting-started Guide" by Albert Bickford (2005), also available at the same web address. This guide is written with sign language researchers in mind, but it is appropriate for all researchers using audio-visual data. The "Getting-started Guide" is short, with little terminology, and enables the creation of an Elan file very quickly. If, however, users already have experience with a different annotation tool for audio-visual data, they can begin with the official Elan manual, which is rather long but well organized and clear. Either way, everyone will need the manual for reference at some stage.

It is quite easy to create an Elan file, to play around with the media controls, and to practice annotating. This is a useful exercise in order to experience some of Elan's capabilities. However, once the concept of an annotation tool is clear and the basic interface of Elan is familiar, users need to spend time considering how they want to annotate their data because the interface must be customized to one's own descriptive needs. The amount of time required depends entirely on how detailed a description researchers want of their data. Each level of description is represented in Elan on a "tier". Tiers can have different relationships to each

other, for example, independent, aligned, or embedded. It takes some time to decide what relationships make sense for the research. However, once annotation has begun, it is not too late to go back and add new tiers (new levels of description). In the most recent version of Elan (2.5.1), one can even change aligned and embedded relationships between tiers. This is a highly valuable function since many researchers develop coding systems in the process of annotation itself, but users should also be aware that the ease with which changes in structural relationships between tiers can be made does not diminish their potential implications for pre-existing annotations.

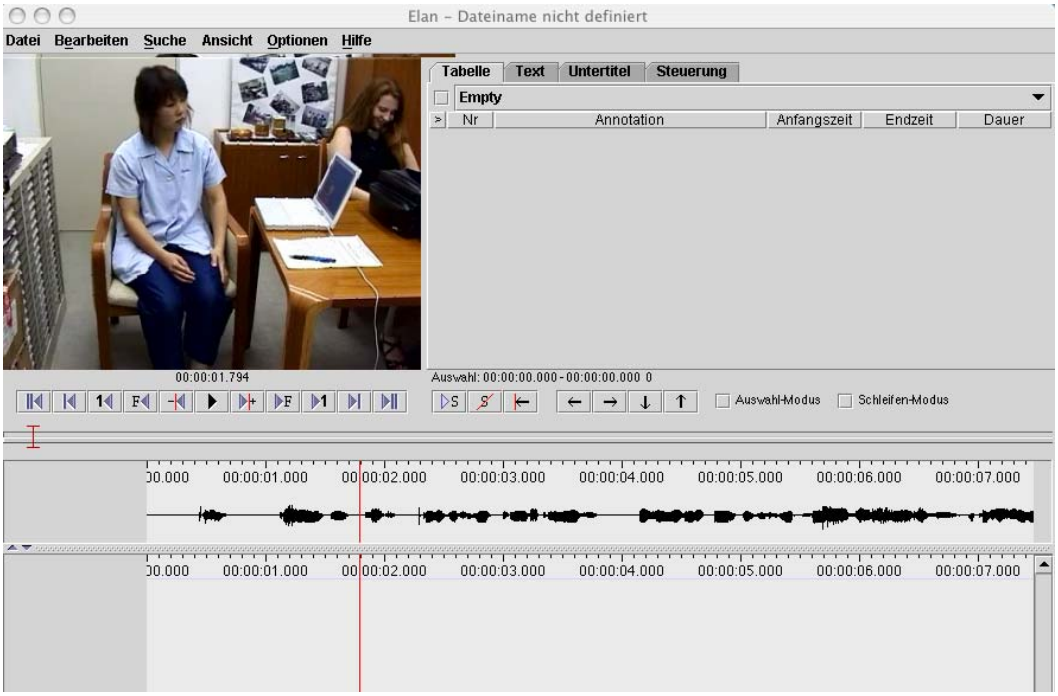

Figure 4: Basic user interface.

Figure 4 shows the basic user interface with no levels of description compared to the customized interface with many researcher-defined levels of description displayed in Figure 5.

As can be seen in Figure 4, the language of the interface can be changed. There are a number of media controls and tools for annotation navigation; a waveform can be used to help with annotation, and speed of playback can be manipulated.

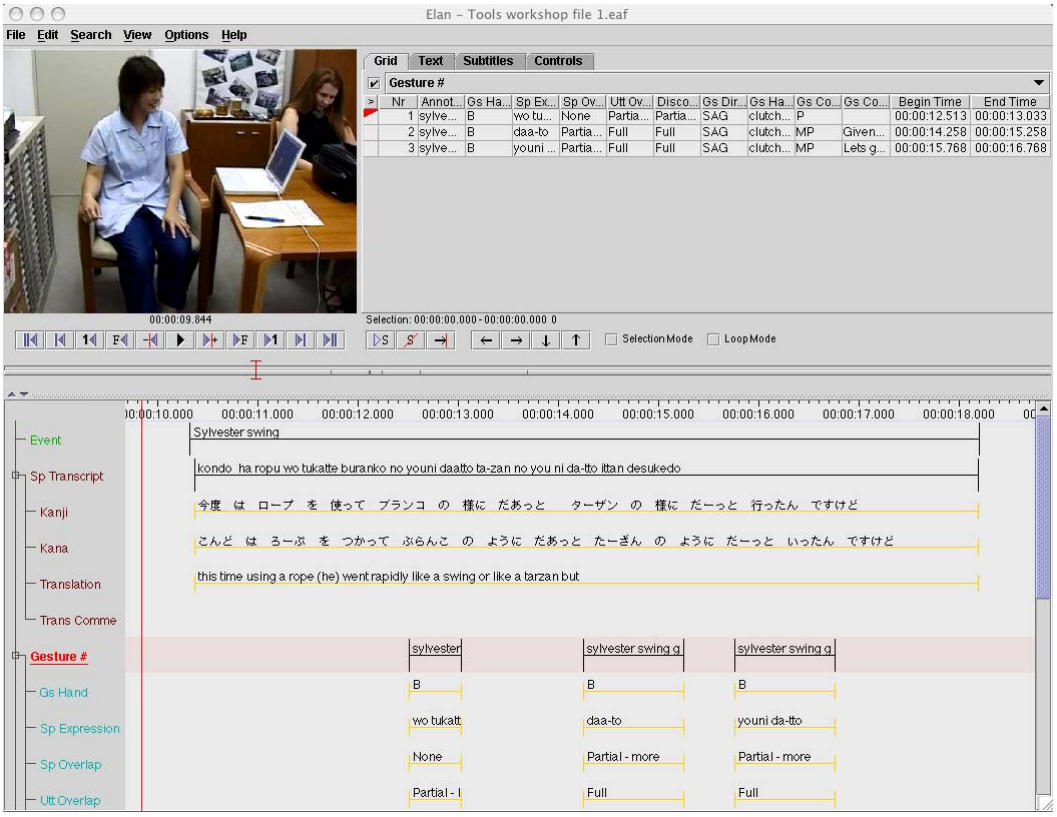

Figure 5: Customized user interface.

The customized interface in Figure 5 shows that annotations are time-aligned, that Elan accepts different character sets, and that relationships between tiers are clearly illustrated.

Data annotation is made easy in Elan in a number of ways. A custom-made Elan file can be stored as a template for use with other media files. Data can be annotated very quickly with use of both the mouse and a number of keyboard shortcuts. Transcriptions can even be imported from other programs, for example Shoebox/Toolbox, Chat, and Transcriber. There are additional productivity enhancements such as semi-automatic segmentation, tokenizing of individual annotations, and tier copying. Errors in annotation can be greatly reduced with a function that allows creation of user-specified vocabulary sets for individual tiers. This also reduces mental load while working with the tool. There is a multiple undo/redo function to enable error correction, and files have an automatic back-up option to protect annotation already done.

To access annotations within a single file, there are versatile search options utilizing sets of constraints. There are also a number of export options for further analyses with other tools. For example, the data shown in the "grid" view in Figure 5 was subsequently exported to a text file with organization maintained in most cases (although the exact procedure depends on the relationships between tiers), imported to a database program such as Excel or Access, and analyzed quantitatively in a statistics program.

3.3.3 Drawbacks

Some limitations of Elan include the search function across multiple files, which is limited to a simple text search without the advantage of user-defined constraints. In addition, annotations on subordinate tiers must occupy the entire duration of annotations on parent tiers. This complicates the viewing of truly relevant subordinate annotations, for example, gestures produced within a single utterance. Like all annotation tools, it is hard to extract portions of the media file along with associated portions of the annotation file for use in presentations, etc. Finally, Elan has slightly less functionality on a Mac, in the "detach media window" option, for example. However, many of these areas are currently under development, and new versions of the tool are released regularly.

3.4 EXMARaLDA: Extensible Markup Language for Discourse Annotation *(Thomas Schmidt)*

3.4.1 Main idea/purpose of the tool

The EXMARaLDA system consists of a data model, a set of corresponding XML formats, and a number of software tools for the creation, management, and analysis of spoken language corpora. EXMARaLDA is developed at the SFB 538 *Mehrsprachigkeit*, a collaborative research center on multilingualism at the University of Hamburg. Its primary objective is to provide a common framework by which the center's projects can share, exchange, reuse, and archive their highly heterogeneous bodies of multilingual data. However, the system's components are made freely available and are also used by a substantial number of students and researchers outside our own institute. Since EXMARaLDA's system architecture, the underlying time-based data model, and the functionality of its tools have been described elsewhere in greater detail (Schmidt 2004, 2005a,b,c), I will limit myself here to a brief summary of the system's most characteristic features and then concentrate on a comparison with other systems covered in this article.

3.4.2 Data model and tools

EXMARaLDA uses a time-based data model that builds on the same idea as the annotation graph (AG) framework proposed by Bird/Liberman (2001), but it is structurally less complex than the general AG formalism. Since the data model is very similar, if not largely identical, to the data models used by such tools as Praat, ELAN, the TASX annotator, or ANVIL, data exchange between EXMARaLDA and these systems is a relatively easy task (well supported by import and export filters in the corresponding tools, see below). In order to enable cross-platform exchange and long-term archivability, EXMARaLDA uses Unicode for the encoding of individual characters and XML files as the primary storage format.

Inputting and outputting EXMARaLDA transcriptions, managing larger bodies of data, and querying corpora for analysis is supported by a number of tools developed in the project. These tools are programmed in Java so that they will run

on all major operating systems (Windows, Macintosh, Linux, and Unix) currently in use. The most important tools are:

1. The EXMARaLDA Partitur-Editor (see Figure 6), a tool for inputting and outputting transcriptions in musical score notation, synchronizing transcriptions with digitized audio or video files, and segmenting ("tokenizing") transcription text into linguistic segments (e.g., words, intonation units, non-phonological material).
2. The EXMARaLDA Corpus Manager, a tool for bundling several transcriptions into a corpus, adding metadata to this corpus, and querying it for the metadata.
3. A query tool (ZECKE) for search across a corpus providing different contextualized views (e.g., a KWIC concordance, a musical score view) of the search result.

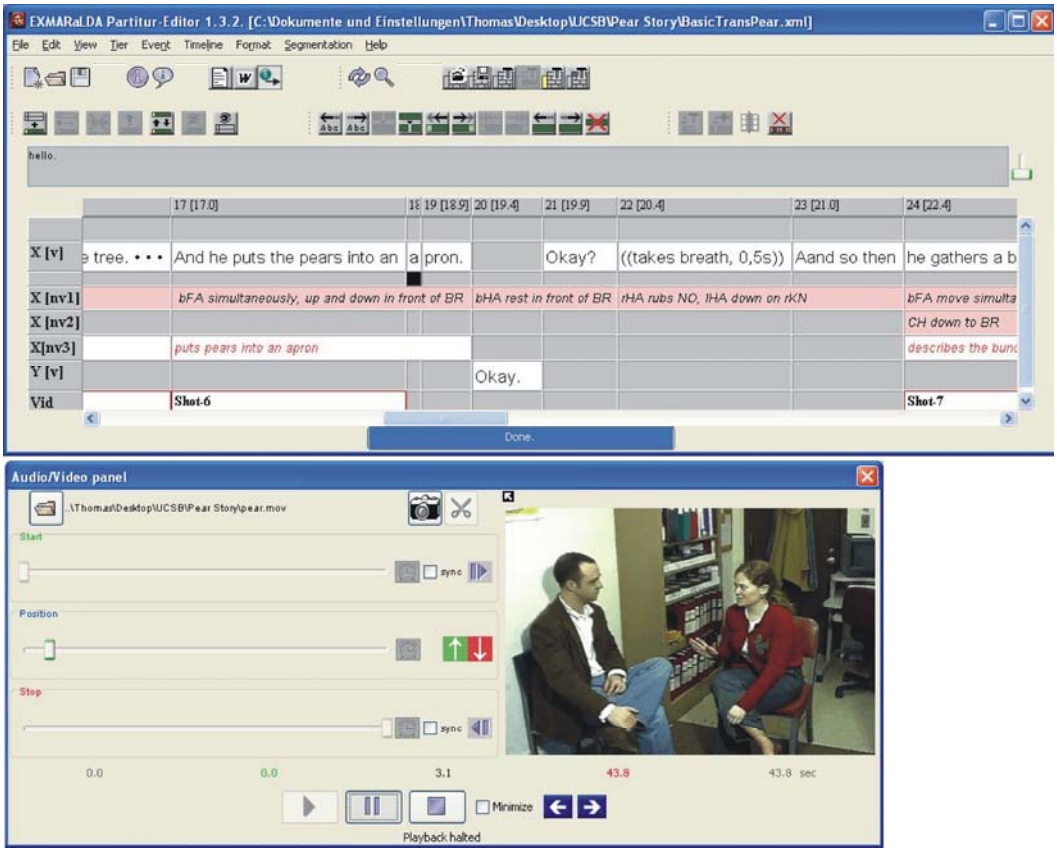

Figure 6: The musical score user interface of the EXMARaLDA Partitur-Editor.

3.4.3 Distinctive features

Interoperability is an important design principle in the development of EXMARaLDA. Many features of the EXMARaLDA tools and data formats are therefore shared by other systems like TASX, ELAN, Praat, and ANVIL. We explicitly encourage users to explore ways of using these and our own tools side-

by-side in order to optimally exploit each tool's strengths and to avoid their weaknesses. That said, we believe that EXMARaLDA's advantages in comparison with the other tools are mainly as follows:

- EXMARaLDA pays much attention to an adequate data visualization for the human user. Most importantly, this includes the possibility to output a transcription on paper in musical score notation (wrapped to a given page width). Other methods make use of the computer's hypertext and hypermedia capabilities to visualize transcriptions on the screen, integrating or linking parts of the transcription to audio, video, or image data. All of these visualization methods are meant to support qualitative, human (as opposed to quantitative, computer-based) analysis of data.
- EXMARaLDA not only regards data exchange with other tools as a potential possibility but also actively supports it through a number of import and export filters integrated into the Partitur-Editor. Data exchange with ELAN, TASX, and Praat is thus possible in both directions. Moreover, the EXMARaLDA Partitur-Editor offers a means of directly accessing Praat's phonetic analysis functions during the transcription process.
- EXMARaLDA directly supports the work with many well-established transcription conventions (HIAT, GAT, DIDA, CHsAT) by providing segmentation algorithms and virtual keyboards for these systems.
- Although EXMARaLDA allows for a close linkage between recording and transcription, it does not require it. EXMARaLDA can therefore be intuitively used also with written language data or when no digitized media file is available.
- EXMARaLDA includes not only a tool for creating and editing individual transcriptions but also tools for subsequent steps like corpus construction, corpus management, and corpus query.

3.4.4 Users and usability

Judging from e-mail feedback to the developers, EXMARaLDA is currently used in teaching as well as in research by several hundred users — mostly in Germany but also in France, Italy, Britain, Switzerland, Austria, and the US. The core user base consists of students and researchers in discourse or conversation analysis and in language acquisition studies, but EXMARaLDA is also employed in pedagogic research, in computational linguistics, and in studies of multi-modality.

Typical EXMARaLDA users are non-expert computer users; that is, their computer literacy does not go much beyond the work with standard office applications. The fact that the software is nevertheless often installed and used without further support leads us to believe that it deserves to be called "user-friendly". Wherever additional support is needed, it is provided through manuals and tutorials on a public website or through individual assistance via a mailing list. EXMARaLDA is designed such that newbies should be able to use its most simple and basic functions after only a short learning phase. It is thus possible to quickly apply EXMARaLDA for ad-hoc or experimental data creation. The more

sophisticated functions will, however, require a more elaborate familiarization with the system's principles. While we do not claim to exempt the prospective user from this task, great care has been taken to support him/her through adequate and publicly available documentation of the system. Since EXMARaLDA has been and is being developed in a process of constant exchange with users from discourse or conversation analysis and language acquisition research, we expect it to be especially intuitive for users from these areas.

3.5 TASX: Time Aligned Signal data eXchange

(Alexandra Thies and Jan Torsten Milde)

3.5.1 Main idea/purpose of the tool

The TASX Annotator (Milde) enables an XML-based annotation of multimodal data on multiple tiers. It was designed to examine "gestural displacement", that is, temporal discrepancies of speech and gesture onsets in L2.

3.5.2 Usability

The screen shot in Figure 7 depicts the two most central components of the TASX Annotator, the video alongside the annotation window, which are time-aligned with each other in order to enable a highly precise location of multimodal data in time. Of particular interest is the annotation window (a graphical user interface, or GUI) with its individual multi-tier set-up, which facilitates a parallel annotation as well as an immediate comparison across the different modalities of interest. The core idea of creating a GUI of this kind was to provide the user with a "virtual sheet of paper," with the striking advantage that – in contrast to an actual sheet of paper – annotations can immediately be loaded into programs such as MS Excel, allowing for temporal calculations, for instance. The GUI itself consists of various components: a set of menus (File, Edit, Tier, Element, Metadata, Options, Tools, Help), a graphical toolbar, which provides direct access to the most prominent functions (open file, save file, load video, load audio, zoom out, zoom in, decrease font, increase font, etc.), as well as a scalable timeline, which informs the user about the currently visible region in the transcript, and a scrollbar at the very bottom, allowing movement from the current region to another one. The dominant area of the TASX GUI, however, is the content area. The content area is used to display and edit transcriptions; hence, this is where most of the interaction with the annotator takes place. Here, the user can switch between three different modes of display: the time-aligned view, the text view, and the table view.

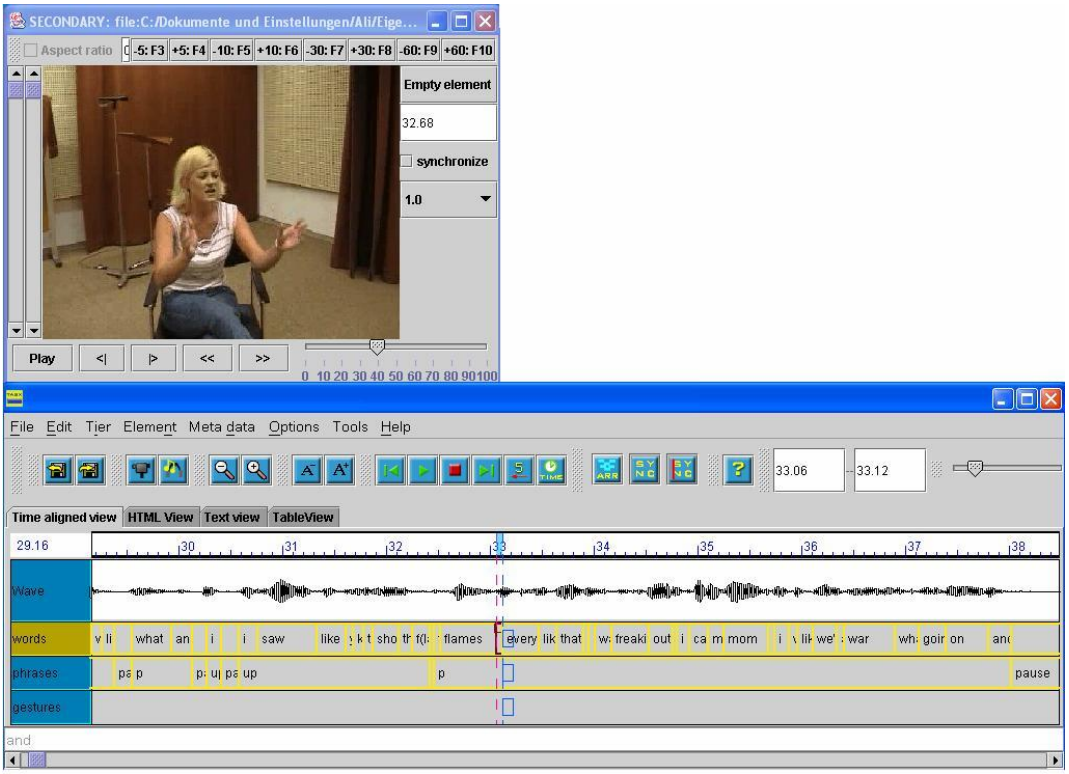

Figure 7: A screen shot of the TASX Annotator showing annotated bilingual language data.

It is the time-aligned view — a two-dimensional grid that is most important during actual annotation. The horizontal dimension is the time axis, while the vertical axis consists of an arbitrary number of annotation tiers (or layers) as well as an optional display of an oscillogram calculated on the basis of the underlying sound file. Each of the tiers consists of a set of separate events (e.g., a word, syllable, or gesture); each of which stores some textual information called a label and is immediately linked to the primary audio/video data by two time stamps, an onset and an offset, which may also overlap with other onsets and offsets. This is even possible on the same tier. In order to create a multi-layered annotation board such as the one depicted in Figure 7, one can either make use of the menu provided at the top of the annotation window or – with ample previous experience with the tool – the keyboard shortcuts. The latter are also indicated in the menu next to each function.

3.5.3 Distinctive features

Aside from the immediate link between transcription and video file as well as the option of playing the video in slow motion, both of which it shares with a great number of comparable annotation tools, one of the core assets of the tool is its rather straightforward usability, allowing even non-specialists to get to grips with it relatively quickly. For more advanced users, the keyboard shortcuts help to significantly speed up the set-up and annotation process. Furthermore, the tool allows for additional fonts to the installed (e.g., IPA, HamNoSys) and for external tools such as Praat or Virtual Dub to be integrated. The tool also lends itself as a

corpus generator, both due to its XML basis and an in-built metadata editor. The search option is especially useful for large extents of annotation. At least theoretically, it is possible to create an unlimited number of tiers; however, despite the option of being able to hide extra tiers, it is questionable whether more than ten tiers can be handled at the same time since the tool lacks a vertical scroll bar.

3.5.4 Drawbacks

The latter leads me to further drawbacks of the tool. Even though the tool allows for the calculation of an oscillogram, the display is far too imprecise for speech annotation – it seems much more practicable to import speech annotations from an external tool such as Praat. While the aforementioned short keys help the user to create a speedier annotation process, certain interferences of label text may occur (the latter of which then unfortunately tend to slow down the process again). Another disadvantage concerning the creation of tiers is that the tiers cannot be hierarchically structured into what one might term "head-tiers" and "sub-tiers". Also, when dragging a segment to another point on the time scale, the video does not move along. The latter would be a useful orientation support for the new placement of the segment. Last but not least, despite the physical presence of an *undo* button in the menu, the function itself has unfortunately been inactive ever since the tool was "born".

As is unfortunately the case with more than a few annotation interfaces that have come and gone over the last 15 years, the development and support of the TASX Annotator seems to have stagnated and, hence, appears a little out-dated in comparison to the other tools covered in this report. It should also be noted that a follow up software is under development: *Eclipse Annotator* (Behrens/Milde 2006). All in all, however, the TASX Annotator is a down-to-earth and relatively easy-to-use tool to annotate and analyze even longer bits of multimodal data.

3.6 MacVisTA: Macintosh Visualization for Situated Temporal Analysis *(Irene Kimbara, Fey Parrill, and Travis Rose)*

3.6.1 Main idea/purpose of the tool

MacVisSTA is a software program developed by VISLab at Virginia Tech to code different aspects of behavior (speech, gaze, gesture, etc.). Its purpose is to allow the user to create time-stamped tags to annotate segments of interaction and to visualize these time-stamped intervals in conjunction with one or more videos. The software works with any QuickTime file and runs on Macintosh OS X. Virginia Tech provides the tool as freeware, available both from the VISLab web page and from SourceForge. The program works best with a fast processor (preferably an Apple Macintosh G5 computer). For coding multi-party conversations, MacVisSTA supports the display of more than one movie file at one time (i.e., synchronized movie files from cameras at different angles). When the coder jumps from one time point to another, these movies (if they are

"genlocked" to begin with) remain synched. Tiers are user-defined and, therefore, quite flexible.

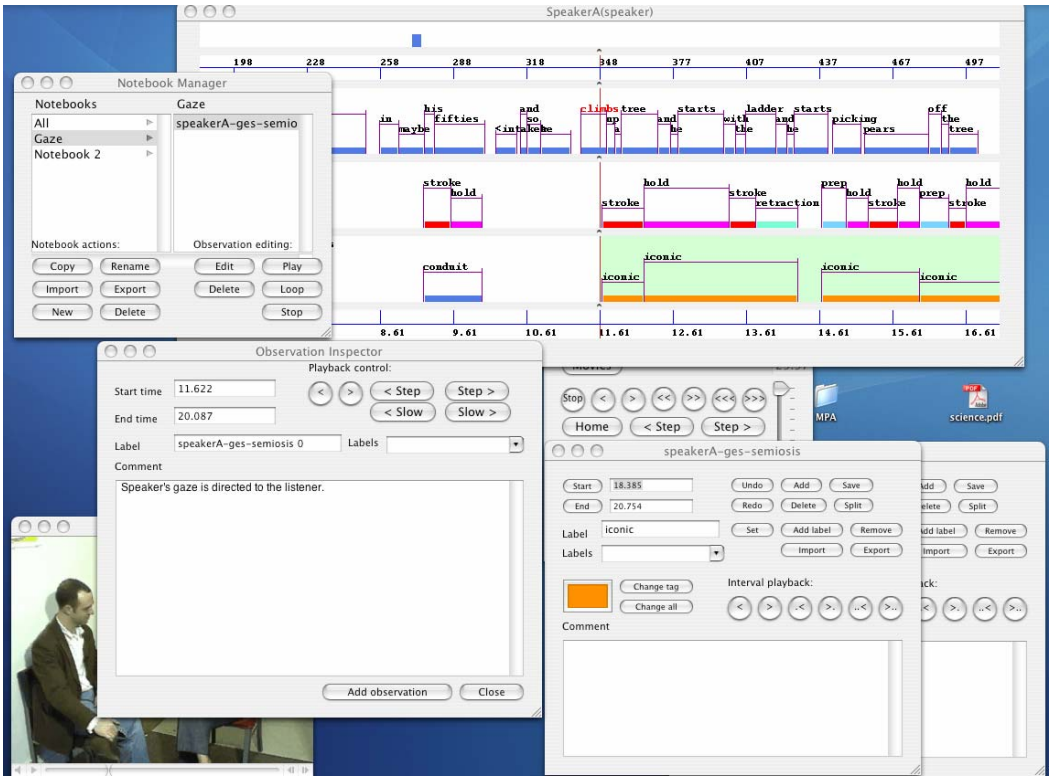

Figure 8: Desktop configuration of MacVisSTA.

3.6.2 Learnability

It is important to note that the tool is still under development. Some of the weakness of the tool can be attributed to this fact. For example, not all of the buttons in windows or functions in menus have been implemented. Since the tool does not come with a complete manual, learning how to use MacVisSTA is not a straightforward process. It takes a great deal of trial, error, and guidance to become a competent user. Novice users can learn best with help from an experienced user. Because of some complexities in the interface, the developer of MacVisSTA is creating a more integrated interface that is more consistent with the user's workflow. Since the tool is still in development, some of its features are subject to change in order to increase its usability, but they will require additional learning.

3.6.3 Speed of execution

Once a project is set up with its associated movie files, user-defined panels, and panes (tiers) in which tags will be created, the speed of coding is not particularly delayed by the design of the tool. Users can select a segment in a tier by dragging the mouse. This lets users watch the movie (without sound) as the mouse is dragged in order to find the behavior of interest. The segment can be played by

pressing the space bar. If the user is satisfied with the location of the segment, it can be made into a tag. There are many keys to set the direction (forward vs. backward) and the speed of playback. Arrow keys also let users navigate through the movie frame by frame. However, there is no sound when the movie is played frame by frame. To have both sound and image, MacVisSTA has an audio control option that allows the user to switch to the audio file (rather than the video file) to navigate through the movie.

Users must create multiple tiers to annotate different aspects of a single behavioral unit (e.g., gesture category, hand used, speaker, etc.). One disadvantage with MacVisSTA is that users cannot create tags with the same begin- and end-points in multiple tiers at the same time. Instead coders need to go to each tier to add a tag and, if necessary, adjust the positioning of the tag by hand. It is also not feasible to display more than about seven tiers at a time because of limited screen space (Figure 8), which can be problematic when annotating multi-party interactions.

3.6.4 Mental load

Annotating behaviors is not a cognitively demanding task, unless the categories used in the annotation are inherently complex (but this cannot be corrected by the software). MacVisSTA includes some useful functions for keeping track of comments. Each tag has three kinds of information associated with it: begin- and end-time, category label, and comment. Comments do not appear in tiers but can be searched later and provide a useful place to note things one has observed about the annotated segment not coded elsewhere. In addition to comments, MacVisSTA has a "notebook" function. Comments and notes are different in how they are organized. Note tags appear in any tier and can span over several tags, making it easy to comment on a series of behavioral units. Notes are kept in a "notebook manager" and users can jump to any note segment by clicking the entries.

Mental load is probably more of an issue when it comes to data analysis. MacVisSTA’s primary function is to create time-stamped labels; thus, little has been done to assist coders with analysis. While tags can be exported as XML files (and sorted and queried in a database), visual analysis of the data is impossible, in part because there is no way to print the MacVisSTA tiers. It is also not possible to open more than one project simultaneously, which makes comparisons across data sets difficult. For example, when comparing two datasets to look for differences in gesture, we (the first two authors) took a series of screenshots, printed them, taped them together, and laid them on the floor side-by-side. This is a laborious process (Figure 9) but may sometimes be a necessary first step to decide what sorts of things to query in a database.

3.6.5 Errors

Setting up a project (defining tiers and linking movies to the project) is relatively error-free. Errors occur frequently during coding. There are some known problems, and some other problems occur sporadically for unknown reasons.

Known problems include associating colors with labels in the pulldown color menu. Sometimes using a different version of the program can solve the problem. Moving projects from one machine to another is also problematic until one becomes familiar with the tool.

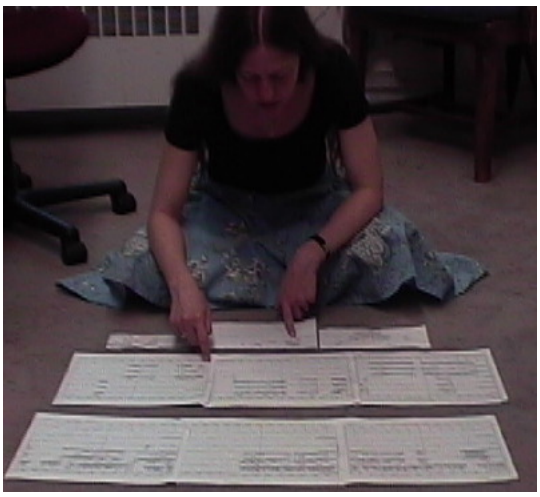

Figure 9: Reviewing corresponding printouts from screen capture.

3.6.6 Other Features

MacVisSTA permits the import of files created with other software such as Praat textgrids. Furthermore, the current version of MacVisSTA can display audio waveforms as well as graphical elements such as motion traces, though we have not yet tested these capabilities.

4 Comparison of the tools

	Media & Text Editors	Anvil	ELAN	EXMARaLDA	TASX	MacVisSTA
Home page	N/A	http://www.dfki.de/~kipp/anvil/	http://www.mpi.nl/tools/elan.html	http://www.rrz.uni-hamburg.de/exmaralda/index-en.html	http://medien.informatik.fh-fulda.de/tasxforce/TASX-annotator	http://vislab.cs.vt.edu/~rtr/
Version used at the workshop		4.5	2.4.1	1.3.2		
Main idea of tool	Video is viewed with video editor; transcriptions typed in text editor with annotation conventions describing non-verbal behavior	Synchronized musicalscore annotation ("partitur") and video	Synchronized multimodal annotation of audio and/or video data	Designed for verbal transcription with non-verbal (e.g., video) seen as an additional "comment". Interface may be less intuitive if verbal and non-verbal behavior are equally important or if the latter is paramount, as timeline maps transcription typography, not absolute duration	Synchronized musicalscore annotation ("partitur") and video	Synchronized musicalscore annotation ("partitur") and video
Technical description						
Supported platforms	N/A	Theoretically runs on any Java-compliant machine, but only known to run well on Windows. Known not to run well on Mac. Said to run under Linux	Java-based, ships on Win, Mac, Linux. Runs well on Windows and Mac. Media playback performance on Mac is slightly less but still satisfactory. Video playback performance is problematic on Linux	Java-based, known to work well on Windows, Linux, and MAC OS 10.x. Sometimes problems with certain media formats on MAC	According to TASX web site, it is Java-based, written and tested on Win98. Also tested (but some features do not work as well) on Win2000, NT, Solaris, Linux, and Mac OS9	Mac OS X 10.2.8 or later. Mac G5 recommended
Supported video formats / codecs	Any used by video editor of choice	Any supported by Java Media Framework v 2.1.1	Depends on the media framework of the underlying operating system. On Windows any format that is supported by Windows Media Player (DirectX) or QuickTime, on Mac OS any format supported by QuickTime, on Linux any format supported by Java Media Framework 2.1.1	Any supported by Java Media Framework v 2.1.1	Any format that is supported by Java Media Framework v 2.1.1	Quicktime
Practical limit on file sizes or durations?	No	Probably not suited for material over 30 minutes	No (at least not known)	No	Tested with video files up to 4 hours. The number of visible elements on a tier is the main limiting factor	It plays 320MB movie file (40min) with no problem; textgrid from Praat slows down the movie and it drops frames

	Media & Text Editors	Anvil	ELAN	EXMARaLDA	TASX	MacVisSTA
Import from:	N/A	Praat	Shoebox, CHAT and Transcriber	Praat, TASX, ELAN, TEI, Simple EXMARaLDA (text file) Also: syncWriter and HIAT-DOS (both require manual post-editing, though)	Praat, Anvil, Transcriber and ESPS Waves+ (using built-in XSLT Proxcessor)	Praat
Export annotations?	N/A	Besides txt, Anvil supports export as tables (separated with tabs e.g.) for statistic software like SPSS and Statistica. This feature is in the Project Tool and a valuable analysis option	Yes: tab-delimited text, "interlinear" text, Shoebox, CHAT, SMIL, QuickTime text track	Yes: TASX, ELAN, Praat, AIF, plain text segment lists, CHAT, TEI, built-in support for "free" conversion with XSL-Stylesheets	HTML, Praat, ESPS Waves+, STM (export via integrated XSLT engine)	Yes, through CocoaMySQL
Annotations saved in XML?	N/A	Yes	Yes	Yes	Binary (more efficient loading) and XML	Yes
Future support for software	Always	Will be maintained for some time to come...	Financial support for new releases for at least several years. Firm support by the institute	At least three more years of funding --> ongoing development, support through e-mail and website	Follow up software Eclispe Annotator is launched at LREC 2006	Uncertain
Open source software?	N/A	No	Yes	No, but I am willing to share code with whoever is interested	Yes	
Cost	Cost of media & text editors	Free	Free	Free	Free	Free
Control modes						
Can view time-aligned annotations with video?	No	Yes	Yes	Yes	Yes	Yes
Handles multiple videos?	Yes	No	Yes, up to 4 videos	No	Yes	Yes, but if they are "genlocked", i.e., video synchronization must be done first, outside of MacVisSTA
Can play video at variable speeds?	Yes	Yes	Yes	No	Yes	Yes
Can move frame-by-frame?	Yes	Yes	Yes. Can move step-by-step, with a minimum step size of 1 millisecond	Yes, if the video codec supports this	Yes (if video format permits)	Yes

	Media & Text Editors	Anvil	ELAN	EXMARaLDA	TASX	MacVisSTA
Can hear audio at slow motion?	Yes (and at frame-by-frame speed)	No. But if audio annotations are imported from Praat, less of an issue	Yes	No. But if audio annotations are imported from Praat, less of an issue	No	Yes, in video primary mode if NOT played frame by frame. Audio is available in frame by frame only in audio primary mode
Search function	Visual and word-processor searches over annotated text	Can search for annotations across multiple files	Complex searches based on temporal and/or structural relations in a single file and simple text search across multiple files	Simple search functionality in the Partitur-Editor itself, more elaborate search functionality in a separate tool (ZECKE, see website)	Rudimentary	No
Analysis functions?			No. Export to tab-delimited text file that can be imported into Excel and similar applications.	Word, utterance etc. segmentation, analysis through XSL-Stylesheets	Yes, using external programs (e.g., Praat)	No
Can code non-verbal behavior in absence of speech?	Yes, although temporal extent must be additionally annotated somehow	Yes	Yes	Yes, but coding **only** non-verbal behaviour (i.e., without at least sporadic verbal behaviour) will not be very intuitive (see above)	Yes	Yes
User-defined annotation types?	Yes	Yes. Done in XML editor outside tool, requires some learning curve	Yes. User can define multiple annotation types based on 4 predefined "stereotypes", within the tool itself and while working on a file	User defined annotation categories (i.e., names for pre-defined types)	Yes. Done inside tool.	Yes. Done inside tool.
Can change annotation definitions after data is annotated?	Yes	You CAN add tracks even for already started annotations. You can change the specification file of the currently loaded annotation and then, reload the file and see if it works. You can also add tags for "ValueSet" attributes. Anvil was developed with the idea of "incremental annotation scheme development" in mind!	New annotation types can always be added. Limited support for changing types of existing annotations	Can easily add new definitions, tiers or speakers	Yes	Yes, color categorization for tags can be easily reassigned. New tiers can also be added later
Can annotations be defined hierarchically?	Yes	Yes	Yes	No	Not supported (though TASX format allows to)	No

Usability (end-evaluation discussed at the workshop)						
	Media & Text Editors	Anvil	ELAN	EXMARaLDA	TASX	MacVisSTA
User-base size	Still quite large	The users on the web list are only a small part of the "real" user base. The mailing list comprises 792 members from 40 different countries and from 318 research institutions (for some institutes you have many users, e.g., MIT, U Edinburgh, U Tokyo, U Chicago, MITRE, and so on)	Unknown; registration is not mandatory. At least several hundreds on mailing lists or otherwise known	Several hundred users in Germany (mostly discourse/conversation analysis and language acquisition research), some users outside Germany	Small (unfortunately)	Regularly used by 1, test-used by 3 in McNeill lab
Positive features discussed at workshop	Ease of learning/training, ease of use, widespread availability, no limit on amount (mins./hrs.) of data that can be processed and analyzed. Robust: system crashes, upgrade incompatibility, inadequate user support, etc. are not issues	See tool description above	Multiple views on the data Multiple undo/redo Use of templates Support for multiple character sets Versatile search options	See tool description above	See tool description above	See tool description above
Other positive features	Not constrained to tabular report formats: annotated transcripts are useful "visualizations" in support of certain analyses; utility for analysis of extended connected discourse sequences	Displays waveform, pitch, and intensity of speech signal.	Displays waveform Up to 4 videos Print function Preliminary 2D video annotation Several automated annotation options	Supports several transcription systems (HIAT, GAT, DIDA, CHAT, IPA) through a number of parameterised functions		
Negative features discussed at workshop	Cumbersome database maintenance and query	See tool description above	No built-in help system More export formats could be added Certain inflexibility of tier structure	See tool description above	See tool description above	See tool description above
Other negative features	No support for inclusion of instrumental analyses of audio or video data, or instrumental measures of nonverbal behavior				User interface is crummy, hard to learn. Despite the available (powerful) extensability features, only very few people committed to the software. Only load a single file, not suited for working with a complete corpus	Under development — many bugs still being fixed

5 Summary

The goal of the workshop described here was not to decide which tool is the best. There is no single best one. In our comparison, it became apparent that tools are only a means to an end. Taking advantage of technology, analysis is supported by tools that have been designed against the background of specific theoretical assumptions (Rohlfing et al., 2005) and for a particular purpose. It depends, therefore, on the researcher's assumptions for her or his analysis which tool is the most appropriate. A good example is assessment of the synchrony between speech and gesture. If synchrony is the primary research issue, a tool has to be chosen that allows for precise objective measurement of this and (in regard to the "mustic score"-interfaces) offers exact time stamps on the time line. Gestural behavior, however, can also be assigned to the words directly, without the time line, if the research issue concerns more the semantics of this relationship.

In our summary, we would like to stress that having so many tools available for multimodal annotations goes hand in hand with the advantage of having a variety of analysis options. However, with regard to rapid technical development and the fact that tools come and go, two points seem to be important:

Firstly, even though all the available tools represent an enormous technical development, a researcher has to be aware of the still valid fact that a hardcopy does not crash and annotations on paper can be universally read. Annotations generated by multimodal annotation tools (excluding the electronic text format) are, to date, easily readable only by a subset of available tools.

Secondly – but related to the purpose of sharing the data – it is important to keep in mind that tools are vulnerable to changing video formats and differently available platforms. One current feature that has the potential to contribute to data exchange between different tools is the export of data in an XML format.

References

Behrens, T. / Milde, J.-T. (2006): The Eclipse Annotator: an extensible system for multimodal corpus creation. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, Italy.

Bird, Steven / Liberman, Mark (2001): A formal framework for linguistic annotation. In: Speech Communication 33(1,2), 23-60.

Kipp, M. (2001): Anvil - A generic annotation tool for multimodal dialogue. In: Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech). Aalborg, 1367-1370.

Kipp, M. (2004): Gesture Generation by Imitation – From human behavior to computer character animation. Boca Raton, Florida: Dissertation.com.

Rohlfing, K. / Bailer-Jones, D. / Loehr, D. / Duncan, S. (2005): How analysis shapes phenomena. Discussion panel at the 2nd Congress of the International Society for Gesture Studies, June 15-18, Université Lyon, France.

Schmidt, Thomas (2004): Transcribing and annotating spoken language with EXMARaLDA In: Proceedings of the LREC-Workshop on XML based richly annotated corpora. Lisbon 2004. Paris: ELRA.

Schmidt, Thomas (2005a): Computergestützte Transkription - Modellierung und Visualisierung gesprochener Sprache mit texttechnologischen Mitteln (Sprache, Sprechen und Computer/Computer Studies in Language and Speech 7). Frankfurt: Peter Lang.

Schmidt, Thomas (2005b): Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech. In: Arbeiten zur Mehrsprachigkeit [Working Papers in Multilingualism], Serie B (62). Hamburg.

Schmidt, Thomas (2005c): Erstellen und Analysieren von Gesprächskorpora mit EXMARaLDA. In: Gesprächsforschung 6, 171-195.

Katharina J. Rohlfing
Universität Bielefeld
Technische Fakultät
Postfach 10 01 31
33501 Bielefeld

Veröffentlicht am 26.7.2006