Emerging Requirements for Multi-Modal Annotation and Analysis Tools

Tony Bigbee, Dan Loehr, Lisa Harper

The MITRE Corporation
{abigbee,loehr,lisah}@mitre.org

Abstract

We review existing capabilities of multi-modal annotation and analysis tools by presenting a survey of seven representative tools, and providing a sample annotation using one system. We discuss emerging requirements including handling electronic ink, eye-gaze tracking, and other time-based considerations. We briefly review aspects of empirically evaluating tool effectiveness and suggest that multimodal interfaces in future analytical tools may be desirable. We conclude by providing a tentative list of desired features for next-generation tools.

1. Introduction

As multi-modal corpora become more prevalent, new annotation and analysis requirements are emerging. Corpora that include time-based data, such as video and marking gestures, make annotation and analysis of language and behavior much more complex than analysis based solely on text corpora and an audio signal. The purpose of this paper is to briefly identify emerging requirements for next generation multi-modal annotation and analysis tools (MAAT) using a survey of current projects as a springboard. Our aim is to focus on areas of annotation and analysis that have received less attention in textual corpora research and that involve multiple levels of temporal phenomena.

To better understand requirements for future analytical systems that support analysis of multi-modal technology and social task settings, we surveyed a sample of current linguistically oriented and behavioral analysis tools. These tools arise from the different traditions of linguistic research and sequential data analysis, and some bridge both traditions. Fisher and Sanderson [1] characterize human computer interaction (HCI) as rich in behavioral, cognitive, and social characteristics; the consequences are that HCI "usually demands questions, data, and methods that defy a single-discipline approach and yield most easily to an exploratory approach that cuts across disciplinary boundaries." They have coined the term "exploratory sequential data analysis" (ESDA) to describe similarities and differences between observational data analysis techniques that use event data.

2. Survey of Existing Systems

There have been several surveys of annotation and analysis tools in the last decade. Sanderson [2] classified and compared 40 different ESDA tools in 1994. The Linguistic Data Consortium website [3] lists and summarizes over 50 tools for linguistic annotation and has recently added a section on gestural analysis. A specific goal of the present survey was to examine how both disciplines' tools integrate *video* data as an initial foray into multi-modality. Our survey consisted of an informal review of available web-based documentation and informal use of software when available. Although there are many excellent tools available, we decided to sample a wide variety. Included tools satisfied one of the following criteria: active development status, "high watermark" or generally recognized as best of breed, or unique capabilities.

Table 1 describes the seven multi-modal annotation and analysis tool projects we examined. All projects but one include video media support. We included MATE because of its unique multi-level architecture, XML integration, and extensibility. Development on some projects, such as MacShapa, has ended. MAAT capabilities are presented in the row headings of table 1. An explanation follows:

- Video: Few tools offer robust video support. Two recent software architectures for managing and playing video are Sun's Java Media Framework (JMF) and Apple's QuickTime. The JMF, for example, does not offer support for the Sorenson codec [4]. No analytical tool offers explicit support for any other time-based media besides audio and video media.

- Music Score: Most tools offer an *interactive music score* scheme for visualizing events in time. Time is mapped from left to right in horizontal space and events at each level of interest are vertically above or below the same horizontal position. This visualization scheme has become nearly ubiquitous. Bird and Liberman [5] depict a visualization of eight annotation levels of a phrase from the Boston University Radio corpus. Because this scheme and some others rely on alphanumeric depictions, it can be difficult to judge patterns over extended time periods.

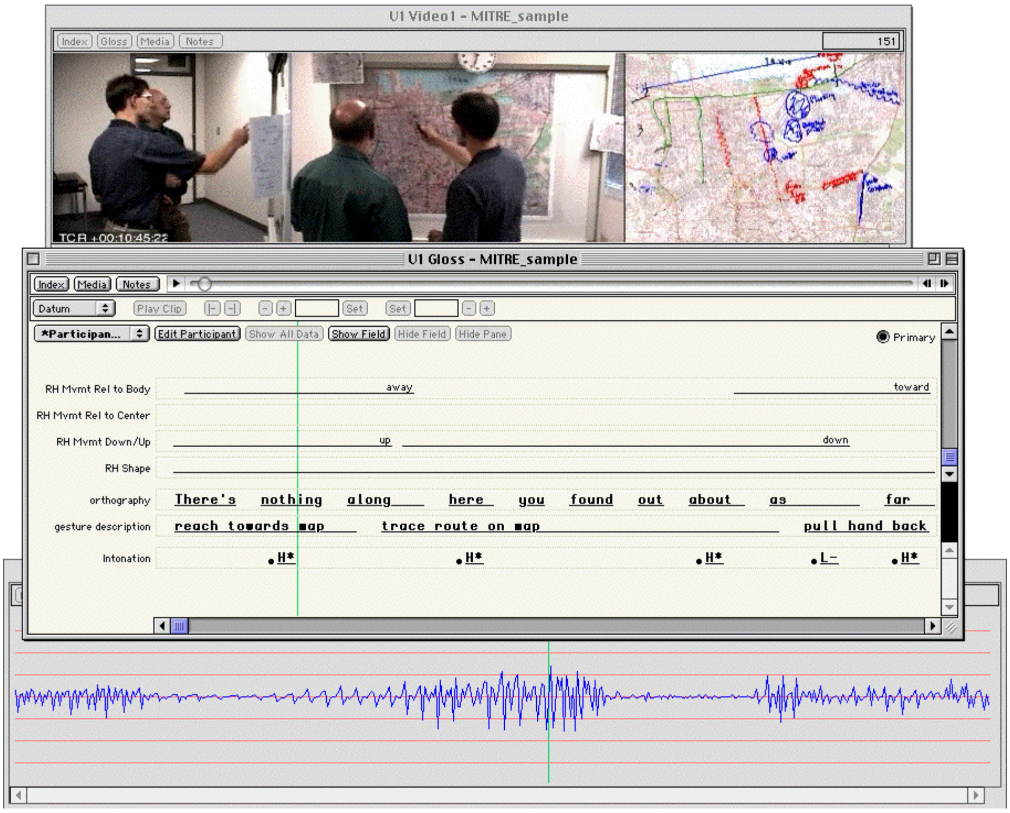
- Multi-level: This term implies the ability to annotate, link between, and analyze different linguistic levels. Levels of analysis may include orthography, morphology, syntax, dialogue acts, co-reference, intonation, gestures, and so forth.

- Analysis: In addition to annotation, some tools provide statistical analysis capabilities. A minimum capability is to search for annotated entities and relationships between them.

As an example of multi-modal annotation, Figure 1 shows a SignStream annotation of two subjects jointly performing a route-planning task. The subjects are marking a plastic-covered paper map taped to an electronic whiteboard.

Table 1: **Survey of Representative Existing Tools**

Product	**Anvil**	**Clan**	**MacShapa**	**MATE**	**MultiTool**	**SignStream**	**syncWRITER**
Version	2.1	2.2	1.1	.17	2.0 beta	2.0	2.0
Evaluation Level	Minimal hands-on	Minimal hands-on	Literature	Minimal hands-on	Minimal hands-on	Medium hands-on	Minimal hands-on
Purpose of Software	Gesture & language annotation & analysis	Child language acquisition	Human behavior annotation & analysis	Multilevel language annotation and analysis	Multi-modal corpora analysis	Sign Language annotation & analysis	Sign Language annotation & analysis
Video	JMF	QuickTime	QuickTime	Audio only	JMF	QuickTime	QuickTime
Import	RST, Praat	Flat text	Unknown	Transcriber, others	TransTool	None	Unknown
Export	Time-stamped XML	Praat	Yes, format(s) unknown	Yes	No	Flat text	Unknown
Music Score	Yes	No	Yes	No	Yes	Yes	Yes
Waveform	No	Yes	No	No	Yes	Yes	No
Multi-level	Yes	No	Yes	Yes	Yes	Yes	Yes
Analysis	No	Yes	Yes	No	No	Yes	Yes
Other Comments	JMF codecs-no Sorenson support.	Supports only 22, 44khz audio.	No download-able user's manual.	XML based annotation. Architecture design.	JMF codecs-no Sorenson support.	Plays multiple sync'd videos.	Output formatting options (visual).
URL	www.dfki.de/ ~kipp/anvil	atila-www.uia.ac.be /childes/	iac.dtic.mil/ hsiac/products /macshapa. html	mate.nis.sdu. dk	www.ling. gu.se/~leifg	www.bu.edu/ asllrp/ SignStream	www.sign-lang. uni-hamburg.de /software/ syncWRITER/info .english.html



RH Mvmt Rel to Body: away toward
RH Mvmt Rel to Center
RH Mvmt Down/Up: up down
RH Shape
orthography: There's nothing along here you found out about as far
gesture description: reach towards map trace route on map pull hand back
Intonation: H* H* H* L- H*

Figure 1: **Sample Multi-Modal Annotation Using SignStream**

The top window of Fig. 1 shows a 3-paned QuickTime video, synchronized and spliced from three different sources together using Final Cut Pro video editing software [6]. The left source is a side camera view. The middle source is a rear camera view. The right source is a view of the map markings, captured electronically by the whiteboard. The bottom window of Fig. 1 shows the sound waveform. The middle window of Fig. 1 contains the time-aligned "music score," showing several types of transcription over a 16 second period.

Here, intonational events such as "H*" (high pitch accent) have been annotated. As SignStream does not do acoustical analysis, we used different software to generate a pitch track, and then manually entered the intonational events into SignStream at the appropriate times. The purpose of this study was to study the relationship between intonation and gesture, and in fact this analysis proved fruitful, discovering complementary discourse functions of the two modalities [7].

The three windows are time-synchronized. The current frame of the video window corresponds temporally to the vertical line in the transcription and waveform windows. One can move the vertical line with the mouse in any window (or scroll through the video with QuickTime controls), and all windows are updated accordingly. This affords a relatively quick, intuitive way to annotate time-based media.

3. Emerging Requirements

We have used this survey to generate ideas on requirements for next generation analytical tools. Linguistic representation, data exchange, and standards have recently received attention (MATE, EAGLES [8], ATLAS [9]). This report focuses on emerging requirements from integrating video data and other time-based media and data. One important advance in video usage is the integration of digital video with annotation and analysis tools. While playing video sequences directly in the display of an analytical system is a major advance over systems that remotely controlled (analog) video decks, this integration presents a host of new opportunities and considerations for designers. Other emerging requirements include handling time-based media other than audio/video, methods for empirically evaluating tools, tool interfaces which are themselves multi-modal, and support for automated aspects of annotation and time tagging individual words from speech

3.1. Time-Based Media Other Than Audio/Video

3.1.1. Ink

In some systems, users may interact with multi-modal interfaces by marking or sketching, where this gestural behavior is significant for the system beyond the fact that a user is making graphical markings on a display. Marking may be used to operate an interface by selecting objects, such as circling, or by quickly sketching out commands that can be interpreted by the system, such as drawing a line through an object to delete it. These markings may be computationally represented in vector form, whereby segments of strokes are explicitly represented along with start and stop times, rather than simply capturing bitmaps of displayed gestures. We have developed software, for example, that converts ink captured electronically from a large whiteboard into Macromedia's Flash vector format, which can also be read by QuickTime players [10]. Other projects, such as Oregon Graduate Institute's QuickSet [11], also represent ink in a vector form, and deliver ink data to multi-modal integration algorithms to be integrated with speech. MITRE and others have also implemented electronic map-based systems in which sketches are non-persistent and visually fade over several seconds.

We do not believe it will be sufficient to present inking gestures or strokes in a video form. We envision direct interaction with ink "objects" whereby users could easily manipulate, copy, move, annotate, and generally interact with captured ink in any part of the annotation and analysis system.

3.1.2. Eye-gaze tracking

Eye movement can be an effective means for investigating visual aspects of attention and the impact of presentation and interaction on user task performance [12]. Eye movements are usually evaluated in terms of fixations—pauses over regions of interest—and saccades, or rapid movements between fixations [12]. Identifying fixation locations is desirable for investigating visual behavior in multi-modal systems usage.

As eye-gaze tracking technology is becoming more robust and affordable, we believe that analytical support requirements will become more pressing. Eye-gaze data will probably require extensive processing after logging as does video. As with ink, eye data use in annotation and analysis tools requires thinking about how to present the two-dimensional spatial characteristics of the data in temporal visualizations.

3.1.3. Other Time-Based Media Considerations

Integrating video, audio, and other time-based media is a challenge because of varying timescales and perceptual characteristics. Dealing with multimedia requires a well-thought out and robust architecture, including time models, media managers, and players. Projects such as MacShapa, SignStream, and DIVA [13] all make use of the QuickTime architecture. Other standards, such as MPEG 4 and MPEG 7, are emerging and may become viable in the future.

An additional consideration is the use of video codecs. Table 2 below illustrates some of the major factors in choosing a codec for analysis purposes based on a 15 minute QuickTime video source file with one uncompressed 16-bit audio track.

Table 2: Example of codec effect

Compression type	File size [reduction]	Video quality
Noncompressed	6.7 gigabytes	Original
Sorenson 2	242 megabytes [4%]	Good
Cinepak	291 megabytes [4%]	Fair

Table 2 must be viewed critically, however, because each compression scheme responds differently to different types of data. Using video compression codecs on graphical data, such as map markings, tends to blur lines/marks and increase file sizes. To keep file sizes small, a smaller viewing size must be used, making marks much more difficult to see. The rightmost panel in Figure 1 is a good example of this problem.

3.2. Empirical Evaluation

We have seen little empirical evidence and evaluation for how the above-mentioned tools support common annotation and analysis tasks. An evaluation strategy that explicitly targets user tasks from usability perspectives should improve user performance and inform the multi-modal community about important issues. Specific factors in designing user interfaces, for example, are: time to learn, speed of performance, rate of errors by users, retention over time, and subjective satisfaction [14]. Discrete, common tasks can be targets for focused evaluation techniques such as controlled experiments. The wide range of inexpensive input devices and emerging visualization techniques offer a suitable basis for choosing variables for experiments. Furthermore, data *presentation* and data *interaction* are user interface aspects that should be investigated in tandem. Each has significant effects on the other during task performance.

3.3. Multi-Modal Interfaces

Interfaces to multi-modal annotation and analysis tools themselves only use modalities available in WIMP (Windows-Icon-Mouse-Pull-down-menus) schemes. Additional modalities for interfacing with a computer are becoming available. Speech input is currently feasible for limited-vocabulary tasks and might be useful for annotation and analysis input tasks. Long et al. [15] have investigated aspects of gesture form for designing interfaces that interpret user gestures. Oviatt [16] has empirically demonstrated benefits of multi-modal interfaces that combine pen and speech input for certain kinds of tasks. Future multi-modal annotation and analysis tools may benefit from multi-modal interfaces by reducing workload in time-consuming, frequent tasks.

4. Conclusion: A Tentative List of Desired Features

In conclusion, we offer a tentative, unordered list of features that we believe are important for next generation systems (Table 3). We generated this list because of our needs in investigating multimodal interfaces and collaborative activity, whether in experiments or field settings like military command and control exercises. No tool we surveyed fully met our needs, and we anticipate other HCI technologies/instrumentation increasing the types and amounts of time-based data available for analysis.

Table 3: Desired features

Video stream(s) time-aligned with annotation	Directly supports XML tagsets
Time-aligned audio waveform display	Acoustic analysis (e.g. pitch tracking) tools included
Direct annotation of video	Hide/view levels
Annotation of different levels	API and/or modular open architecture
Music-score display	Automatic tagging facilities
Easy to navigate and mark start and stop frame of any video or audio segment	User can select current audio track from multiple available audio tracks
Segment start and stop points include absolute time values (e.g. not just frames)	User can create explicit relationships or links across levels
Can specify levels and elements (attribute / values)	Inclusion of graphics as an annotation level
Support for overlapping, embedding and hierarchical structures in annotation	Easy to annotate metadata (annotator, date, time, etc.) at any given level or segment
Some levels time-aligned, others are independent but aligned in terms of segment start / stop times	Support for working with multiple synchronized video, audio, and vector ink media sources
Import/export all annotations	Cross platform execution
Query/search annotations	

5. References

[1] Sanderson, P. & Fisher, C., "Exploratory Sequential Data Analysis: Foundations", *Human-Computer Interaction*, 9 (3): 251-317, 1994.

[2] Sanderson, P.M. *Exploratory sequential data analysis: software*, Technical Report EPRL9401, University of Illinois at Urbana-Champaign, Engineering Psychology Research Laboratory, Department of Mechanical and Industrial Engineering. 1994.

[3] <http://morph.ldc.upenn.edu/annotation/>

[4] <http://java.sun.com/products/java-media/jmf/2.1.1/formats.html>

[5] Bird, S. & Liberman, M., "A formal framework for linguistic annotation", *Speech Communication*, 33(1,2), 23-60, 2001.

[6] <http://www.apple.com/finalcutpro/>

[7] Loehr, Dan. *Intonation, Gesture, and Discourse*. Proceedings, Georgetown University Round Table on Languages and Linguistics, 2001.

[8] Leech, G, Weisser, M., Wilson, A., & Grice, M.*, Survey and guidelines for the representation and annotation of dialogue*, 1998. <http://www.ling.lancs.ac.uk/eagles/delivera/wp4final.htm>

[9] Bird, S., Day, D., Garofolo, J., Henderson, J., Laprun, C. Liberman, M., "ATLAS: A Flexible and Extensible Architecture for Linguistic Annotation"*, Proceedings of the Second International Language Resources and Evaluation Conference*, Paris: European Language Resources Association, 1699-1706, 2000.

[10] <http://www.mitre.org/tech_transfer/softboard_conversion/index.htm>

[11] Cohen, P. R., Johnston, et al., "QuickSet: Multimodal interaction for simulation set-up and control"*, Proceedings, Fifth Conference on Applied Natural Language Processing*, Washington, D.C., 1997.

[12] Salvucci, D. D., & Anderson, J. R., "Automated eye-movement protocol analysis", *Human-Computer Interaction*, in press.

[13] Mackay, W. & Beaudouin-Lafon, M. , "DIVA: Exploratory Data Analysis with Multimedia Streams", *Proceedings of ACM CHI '98 Human Factors in Computing Systems*, Los Angeles, California, ACM/SIGCHI, 1998.

[14] Shneiderman, Ben. *Designing the User Interface*, 3rd ed. Reading, MA: Addison-Wesley Publishing Co., 1998.

[15] Long, A. C., J. A. Landay, et al., "Visual Similarity of Pen Gestures", *Proceedings of Human Factors in Computer Systems* (SIGCHI), 2000.

[16] Oviatt, S.L., DeAngeli, A., & Kuhn, K., "Integration and synchronization of input modes during multimodal human-computer interaction." *Proceedings of Conference on Human Factors in Computing Systems*, CHI '97, ACM Press, 1997.