# Automated Support for AARs: Exploiting Communication to Assess Team Performance

Noelle LaVoie
Parallel Consulting
Longmont, Colorado
lavoie@parallel-consulting.com

Peter Foltz, Mark Rosenstein
Pearson Knowledge Technologies
Boulder, Colorado
pfoltz@pearsonkt.com,mbrmbr@acm.org

Rob Oberbreckling
Perceptive Research
Boulder, Colorado
rob.oberbreckling@gmail.com

Ralph Chatham
ARPA Consulting
Falls Church, Virginia
ralph.chatham@verizon.net

Joe Psotka
U.S. Army Research Institute
Arlington, VA
joseph.psotka@us.army.mil

## ABSTRACT

The After Action Review (AAR) process provides a powerful methodology that in the context of training maximizes the benefits of exercises by enabling a unit to learn from experience by systematically reflecting on their strengths and weaknesses. We have developed a tool that supports the AAR process, essentially extending an Observer Controller's (O/C) reach automatically. This tool was developed with two training contexts in mind: live STX lane convoy training at the National Training Center (NTC) and simulated convoy training using DARWARS Ambush! at the Mission Support Training Facility at Fort Lewis. At NTC, live radio communication is captured during training, while with Ambush! communication using voice over IP (VOIP) is recorded. The tool automatically converts recorded speech to text and then analyzes the text, using advanced statistical machine learning technologies, to determine a unit's performance and identify critical incidents, leading indicators, and other training events that could be included in an AAR.

We worked closely with Subject Matter Experts (SMEs) to derive the important dimensions of performance allowing the tool to support a wide range of O/C and commander AARs. The tool rates a unit on several scales based on a mission essential task list (METL), including command and control, situation understanding, use of standard operating procedures (SOPs), and battle drills. For each rating scale, the tool selects appropriate training events that reflect the unit's range of performance from untrained through practiced to trained. The tool's interface makes it easy to spot performance weaknesses at a glance and then to drill down to understand these weaknesses by listening to the relevant radio communication. The tool also enables commanders to create a custom AAR by selecting events of interest and the associated radio communication and then adding their own comments.

## ABOUT THE AUTHORS

**Noelle LaVoie** is a founder of Parallel Consulting, LLC where she acts as the lead Cognitive Psychologist. Parallel Consulting specializes in combining qualitative and quantitative methodologies in conducting applied social science research. Previously Noelle held the position of Senior Member of Technical Staff at Pearson Knowledge Technologies, where she focused on developing innovative applications of Latent Semantic Analysis (LSA) and other machine learning technologies. These included military applications involving tacit knowledge based assessment of military leadership, online collaborative learning, visualization tools to support multinational collaboration and design of interactive electronic manuals. Noelle received her Ph.D. in Cognitive Psychology from the University of Colorado, Boulder, in 2001. Contact information: Parallel Consulting, 806 Bowen Street, Longmont, CO 80501, lavoie@parallel-consulting.com.

**Peter W. Foltz**, Ph.D. is founder and Vice President for Research at Pearson Knowledge Technologies and Senior Research Associate at the University of Colorado, Institute of Cognitive Science. His research has focused on computational modeling of knowledge, team research, and technologies for automated training assessment. He has published a range of articles on Team assessment, information retrieval, natural language processing, training technology, clinical diagnosis, and cognitive modeling. Peter has served as principle investigator for research for

the Army, Air Force, Navy, DARPA, National Science Foundation, and Intelligence Agencies. Contact information: Pearson Knowledge Technologies. 4940 Pearl East Circle, Suite 200, Boulder, CO, 80305. pfoltz@pearsonkt.com.

**Mark Rosenstein** is a Senior Member of Technical Staff at Pearson Knowledge Technologies applying machine learning and natural language processing techniques to problems involving understanding and assessing language and the activities connected with the use of language. Contact information: Pearson Knowledge Technologies. 4940 Pearl East Circle, Suite 200, Boulder, CO, 80305. mbrmbr@acm.org.

**Rob Oberbreckling** is a founder of Perceptive Research Inc. His interests include applying software systems to problems in cognitive science, natural language processing, machine learning, audio signal processing, and automated human performance measurement, modeling, and assessment. Robert previously was a Senior Member of Technical Staff at Pearson Knowledge Technologies where he led team communication instrumentation and data collection efforts in the field as well as created predictive systems for individual and team performance for commercial and military applications. Contact information, Perceptive Research Inc., 3050 24th St., Boulder, CO 80304, rob.oberbreckling@gmail.com.

**Ralph Chatham** is a physicist, storyteller, all-purpose curmudgeon, and lately program manager for the Defense Advanced Research Projects Agency. He is currently a private insultant, delivering advice on technology development, and training in the Defense Department. He has been a submarine officer, laser builder and chairman of two task forces of the Defense Science Board, herding DoD elephants to explore the issues of training superiority and training surprise. He has managed, either inside or outside the government, contract research on: putting lasers in space to talk to submarines patrolling under water and clouds; synthetic aperture sonar; real science applied to detecting deception; and digital tools, games and simulations for training such things as language and information technology troubleshooting. He created and managed from afar the research program discussed in this paper. In addition to the Defense Superior Service Medal, the Secretary of Defense Medal for Exceptional Public Service and other DoD award, Ralph and his wife jointly received a 2003 National Storytelling Network Oracle Award. Contact Coordinates: 2631 Kirklyn Street, Falls Church, VA 22043; 703 698 5456; ralph.chatham@verizon.net.

**Joseph Psotka** is a Program Manager for basic and applied research in behavioral and social sciences at the Army Research Institute. He earned a Ph.D. degree in cognitive psychology from Yale University in 1975. He taught at several colleges and universities, including Southern Connecticut State College and the University of Waterloo, before becoming Director of Research at NPSRI in Alexandria, Va. in 1978. He was made a Resident Scholar of the National Institute of Education (NIE) in 1981. Dr. Psotka joined the Army Research Institute in 1982 as a team chief within the Training Laboratory, where he has remained. In 1988 his edited volume on Intelligent Tutoring Systems: Lessons Learned was published. Recently, he has edited a special edition of "Intelligent Learning Environments" on the use of LSA for instruction in threaded discussions. His research now focuses on social network analysis, LSA and automated text understanding, leadership, communities of practice, unobtrusive measurement technologies, automated tutoring by intelligent agents, simulation technologies, and higher order thinking. Contact information:
Joseph Psotka, US ARI, 2511 Jefferson Davis Highway, Arlington, VA 22202-3926. joseph.psotka@hqda.army.mil

# Automated Support for AARs: Exploiting Communication to Assess Team Performance

Noelle LaVoie
Parallel Consulting
Longmont, Colorado
lavoie@parallel-consulting.com

Peter Foltz, Mark Rosenstein
Pearson Knowledge Technologies
Boulder, Colorado
pfoltz@pearsonkt.com,mbrmbr@acm.org

Rob Oberbreckling
Perceptive Research
Boulder, Colorado
rob.oberbreckling@gmail.com

Ralph Chatham
ARPA Consulting
Falls Church, Virginia
ralph.chatham@verizon.net

Joe Psotka
U.S. Army Research Institute
Arlington, VA
joseph.psotka@us.army.mil

A cornerstone of the Army training system is the After Action Review (AAR), (Morrison & Meliza, 1999) which is conducted after every training or operational mission. The AAR process provides a powerful method for maximizing the benefits of training exercises by enabling a unit to learn from experience.  The key is for the AAR leader, often the unit commander, to systematically review the unit's performance and provide feedback on their strengths and weaknesses.  In the context of training at the NTC, Observer/Controllers (O/Cs) often facilitate learning by providing much of the feedback to the unit directly. A similar arrangement is often used for units training with Ambush! in which the unit commander works with the O/Cs running the simulation to provide detailed feedback to the unit.

In order to conduct an AAR the O/C must provide specific incidents and interactions from training. STX lane training at NTC often occurs over several hours, increasing the AAR leader's workload, and making it more difficult for the unit commander or O/C to produce specific examples from training.  We have developed a tool set that supports the AAR process by essentially extending an O/C's reach automatically. This toolset is built upon technologies that enable automatic monitoring of team and individual performance through analyses of their communications. This paper will describe the technological approach and the development of the AAR toolset.

## AUTOMATED COMMUNICATION ANALYSIS

Verbal communication provides a rich source of information about a team's performance, including what team members know, how information flows through the team's network, and detailed information about cognitive states, situation awareness, workload and stress. In fact, within the distributed training community, trainers and subject matter experts typically rely on listening to a team's communication in order to assess that team's performance. In order to exploit the information inherent in verbal communication, technologies are needed that can assess both the content and patterns of the verbal information flowing in the network and then convert the analyses into straightforward, usable feedback for teams and commanders.

The overall goal of automated verbal communication analysis is to apply a set of computational modeling approaches to networked communication in order to convert the verbal communication into useful characterizations of performance.  These characterizations include metrics of team performance, feedback to commanders, and alerts about critical incidents related to performance.  This type of analysis has several prerequisites.  The first is the availability of sources of clear verbal communication.  Second, there must be performance measures which can be used to associate the communication to actual team performance.  Finally, these prerequisites can be combined with computational approaches applied to the communication in order to perform the analysis.  These computational approaches include computational linguistics methods to analyze communication, machine-learning techniques to associate communication to performance measures, and finally cognitive and task modeling techniques.

By applying the computational approaches to the communication, we have a complete communication analysis pipeline. Communications are converted directly into performance metrics which can then be incorporated into visualization tools to provide commanders and Soldiers with applications such as automatically augmented AARs and debriefings.

A number of AI, statistical, and machine learning techniques have been applied to discourse modeling, generally for the purpose of improving speech recognition and dialogue systems. However, few have focused directly on the content of team discourse. Recent methods that have been tested include decision trees (Core, 1998), statistical modeling based on current utterance and discourse history (Chu-Carroll, 1998), and hidden Markov models. For example, Stolcke et al., (2000) were able to predict the tags assigned to discourse within 15% of the accuracy of trained human annotators, while Kiekel et al., (2004) developed Markov models of communication patterns among team members that were able to predict overall performance.

Some of these components have been previously developed and an earlier version was successfully evaluated, demonstrating that the toolset effectively predicts aspects of objective mission performance by measuring the quality of verbal team interactions. The ability to produce a team assessment and monitoring system is made possible by a technology for mimicking human understanding of the meaning of natural language. The basic technology is a machine learning method called Latent Semantic Analysis (LSA). LSA is a fully automatic corpus-based statistical modeling method for extracting and inferring relations of expected contextual usage of words in discourse (Landauer, Foltz & Laham, 1998). In LSA training texts are represented as a matrix, where each row represents a unique word and each column represents a text passage or other unit of context. The entries in this matrix are the frequencies of the word (rows) in the context (columns). A singular value decomposition (SVD) of the matrix results in a 100-500 dimension "semantic space" where the original words and passages are represented as vectors. The meaning of any passage is the sum of the vectors of the words in the passage (Landauer et al., 1997). Words, utterances, and whole documents can then be compared against each other by computing the cosine between the vectors representing the texts. This technique provides a measure of the semantic similarity of two texts, even if they do not contain words in common. LSA has been used for a wide range of applications and for simulating knowledge representation, discourse and psycholinguistic phenomena. These approaches have included: information retrieval (Deerwester et al., 1990), automated essay scoring (Landauer et al., 2000), automated text analysis (Foltz, 1996), and have been incorporated into a number of commercial text processing applications, such as Apple Computer's spam detection.

Initial tests using LSA for team communication analysis have shown great promise. Typically, LSA is first automatically trained on a body of text containing knowledge of a domain, for example a set of training manuals related to the tasks from which the communication is drawn. After such training, LSA is able to measure the similarity of meaning of two utterances in a way that closely mimics human judgments. Using existing communication data, the technology is able to provide accurate predictions of overall team performance, make reliable judgments of the types of statements each team member is making, and predict team performance problems based on the patterns of communication among team members.

Over a series of studies, LSA-based communications methods have been evaluated favorably in terms of their ability to predict team performance. For instance, LSA was successfully able to predict team performance scores in simulated task environments based only on communications transcripts (Foltz, 2005; Foltz, Martin, Abdelali, Rosenstein & Oberbreckling, 2006; Gorman, Foltz, Kiekel, Martin & Cooke, 2003; Kiekel, Cooke, Foltz, Gorman & Martin, 2002). Using human and automatic speech recognition system (ASR) transcripts of team missions in a UAV environment, in simulators of F-16 missions, and in Navy TADMUS exercises, LSA predicted both objective team performance scores and SME ratings of performance at very high levels of reliability. These results illustrate that LSA-based methods can successfully determine the overall performance of a team based on their verbal communications.

Because team communication is typically spoken, ASR can be applied to convert speech to text for input into the toolset. LSA has been tested for the analysis of ASR input for a limited portion of a dataset of verbal communication. The results indicated that even with typical ASR systems degrading word recognition by 40%, LSA's prediction performance degraded less than 10% (see Laham, Bennett & Derr, 2002 and Foltz, Laham & Derr, 2003). Note that because verbal interactions in such situations are highly constrained by the actions currently being taken and by the current execution status of the mission plans, and are largely routinized, the difficulties of both automatic speech recognition and LSA understanding are greatly reduced. Moreover, because LSA derives meaning from whole utterances, not from individual words, it is immune to fairly high word level error rates typically found in speech recognition systems.

The present work sought to expand the automated communication analysis results by incorporating additional statistical language modeling techniques in conjunction with LSA. The goal was to then develop a toolset to support automated AARs based on these modeling techniques.

## DATA COLLECTION

Two datasets were collected and analyzed during this effort. In collaboration with the Fort Lewis Mission Support Training Facility, we collected audio, video and event log data from the DARWARS Ambush! virtual environment convoy training. In Ambush! up to 50 Soldiers jointly practice battle drills and leadership during simulated convoy operations. At the National Training Center (NTC), Fort Irwin, a second dataset was collected consisting of data from live mounted convoy STX lane training. In collaboration with the NTC Observer/Controllers (O/Cs) performance assessments of the datasets and recorded AARs and hot washes from the live training exercises were collected. Both data collection efforts concentrated on platoon and squad-level teams performing convoy operations.

Both in Ambush! and at NTC units are trained in situations currently encountered on a daily basis by units deployed for Operations Enduring Freedom and Iraqi Freedom. In the training, company-sized elements receive a fragmentary order (FRAGO) to conduct a mounted tactical patrol along a specified route. The convoy commander conducts troop-leading procedures, issues a movement order, and leads the convoy along the designated route. The convoy encounters contacts along the route, which can include a civil disturbance, a rocket-propelled grenade attack, an improvised explosive device (IED), a near ambush, vehicle-borne IED (VBIED), negotiation with Iraqi police and complex attacks (IED and ambush) (see Kuhn, 2004).

### DARWARS Ambush!

DARWARS Ambush! is a widely used game-based training system that has been integrated into training for many brigades prior to deployment in Iraq (Diller, Roberts, Blankenship & Nielson, 2004; Diller, Roberts & Wilmuth, 2005). In this environment up to 50 Soldiers are able to jointly practice battle drills and leadership training during simulated convoy operations. Figure 1 shows a typical user's view during training.

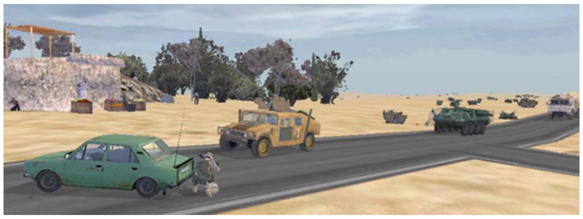


**Figure 1. DARWARS Ambush! training scenario screen.**

At Fort Lewis, we were able to coordinate the collection of over twenty-two DVDs containing over 250 training missions of approximately a half hour apiece including VOIP audio communication, and video and event logs in some cases.

### National Training Center

Data collection at the NTC was significantly more challenging than collection of the Ambush! data, as might be expected from trying to instrument real platoons and squads in the field. We collected voice activated recordings of SINCGARS FM communications during STX lane training, although topographical features made FM signals unavailable or degraded from several locations, so voice quality was not as high as in the controlled Ambush! environment.

Data was collected during rotations from January through June of 2007. We recorded a total of 105 STX lane training missions, of which we selected 57 recordings that had acceptable quality audio, and training events of interest. These recordings varied in duration from as little as ten minutes to several hours. Combined with the 250 missions recorded from Ambush! at Fort Lewis, we collected a total of over 300 training missions.

## PERFORMANCE METRIC DEVELOPMENT

Providing feedback on team performance requires the toolset to associate performance metrics with communication streams. Thus, in addition to the audio communication, the system typically requires one or more metrics of team performance. There are a wide range of issues in determining appropriate metrics for measuring team performance (e.g., Brannick, Salas, & Prince, 1997). For example, metrics need to be associated with key outcomes or processes related to the team's tasks, they should indicate and provide feedback on deficiencies for individuals and/or teams, and they need to be sufficiently reliable so that experts can agree on both the value of the metric and on how it should be scored for different teams (Paris, Salas & Cannon-Bowers, 2001).

Performance metrics can include objective measures of performance, such as threat eliminations or mission objectives completed, or subjective measures of performance, such as Subject Matter Experts' (SME) ratings of aspects of performance including command and control and situation awareness. Additionally, components of evaluations made during AARs, such as identification of specific critical incidents, failures, or errors can be used to measure performance. In order to be effective, the human performance measures must capture important aspects of team performance for assessing proficiency, as well as changes in performance that signal critical events and potential trouble. In both the Ambush! and NTC convoy training contexts, evaluation occurred as part of the AAR process, so it was important that the performance measures were drawn from the same task context, and developed in conjunction with SMEs having extensive experience working with convoys. In short, the primary goal of these measures was to provide a stable, reliable and valid indication of human performance based solely on communications collected during convoy training operations.

The best way to develop performance measures is to rely on SMEs who have extensive experience working with convoys and are able to translate that experience into evaluations of units' performance. In addition to selecting good SMEs, it is crucial to develop appropriate rating scales that accurately capture the important dimensions of team performance over the course of a training mission. The first step in developing rating scales was to understand the tasks involved in conducting convoys. This included understanding the range of performance, the types of information available through communication, and the feedback that commanders and trainers typically provide to units during training to improve performance. This was accomplished by observing convoy training conducted using the Ambush! simulation at Fort Lewis and during STX lanes training at the National Training Center. Right seat ride-alongs with Observer/Controllers at the NTC and interviews with SMEs were carried out to better understand which aspects of performance experts use to assess a unit's proficiency and to help identify the appropriate level of analysis to code convoy performance. During the interviews SMEs were asked to listen to audio clips of convoy training sampled from the collected data and describe the performance of the unit to allow us to better understand the features of performance that were available in the audio communication alone.

It became clear early on that audio communication is an extremely rich source of information for the SMEs, and that most SMEs were able to evaluate how well a unit was performing within the first few minutes of an audio clip. One SME was identified as being a particularly reflective practitioner: he was able to clearly articulate the important aspects of performance observable in the communication and explain how these changed over the course of a mission. We decided to work closely with this SME, LTC (Ret) Cyle Fena. He acted as our primary SME, guiding the development of the rating scales used by the other SMEs to rate the audio communication. At the time this work was conducted LTC Fena was the branch manager for Echelons Above Brigade Security, Plans and Operations at Fort Irwin, working as a contractor for Northrop Grumman.

### Rating Scales

To increase both validity and reliability of the ratings it was extremely important that the scales we developed were relevant and meaningful to the SMEs rating the communication. It was decided to base the scales on Army doctrine, and in particular on the concept of Mission Essential Task Lists (METL), (see FM 3-0, Army Operations and FM 7-1, Battle Focused Training). Using METL, we developed four scales that captured the important dimensions of performance in this domain: command and control, situation understanding, adherence to standard operating procedures and battle drills. We also included a general Team Performance scale, which in previous research has been a good predictor of a unit's proficiency (Foltz, Martin, Abdelali, Rosenstein & Oberbreckling, 2006). The Army's standard three point rating scale of Trained, Practiced, and Untrained was expanded into a five point scale anchored at the top (Trained), middle (Practiced) and bottom (Untrained). These were the scales and ratings used by the SMEs to evaluate events.

### Rating Tool

A rating tool was designed to support SME ratings of the audio communication. The tool allowed the SMEs to listen to the audio, select audio segments, and then rate the segments. The rating tool presents the audio in a visual format that allows a user to interact directly with audio while it is being played. By clicking and selecting a segment of audio, the segment is marked as a training event of interest and the SME is prompted to rate the segment. One benefit of this tool is that it automatically captures the event, the corresponding audio, and the SME generated rating data. The ease of interaction with the tool facilitated testing various rating scales and assessing their value by making it easier and faster to try various rating schemes and then examine the ratings produced.

Seven SMEs rated the collected audio on these scales, using the rating tool. The SMEs listened to the mission audio and marked training events. They then rated the events using the scales we developed. The SMEs were also asked to distinguish between critical events, defined as events that change the scope of battle, the commander's plan or disrupt the operational tempo, and other training events in the communication.

Finally, SMEs conducted AARs for every mission they rated, providing sustains, improves, and overall ratings for each mission.

### Rating Reliability

Before using SME ratings as a performance measure, it is critical to assess how well the SMEs agreed with each other. All SMEs were asked to provide ratings for a pair of missions selected for the purpose of computing reliability and agreement. Intraclass correlations among the SMEs ranged from .76 to .85 (p<.001) for average items suggesting excellent reliability. Exact agreement was calculated between every pair of SMEs, and average exact agreement ranged from 24% to 50%. Average adjacent agreement, which includes the ratings within one point, ranged from 74% to 96%. Two SMEs had extremely high agreement, with their adjacent agreement ranging from 93% to 100%, and exact agreement ranging from 51% to 86%. The agreement among SMEs was impressive, and indicates that the SME ratings are appropriate for computational modeling. It also provides support for the prerequisite that SMEs are able to reliably detect performance from communication.

## DATA MODELING

To go from audio data and SME ratings to a system that can automatically rate new missions requires building predictive models of the data. The goals of modeling were to identify critical events in segments of audio communication and assess team performance to provide feedback to Soldiers and support automated AARs. Data modeling was conducted on a set of 72 training missions which included communication data, speech analysis variables, and SME-selected critical events and ratings of performance.

Critical event modeling was conducted using a spectrum method utilizing discrete time windows where the size of the window, and step size between windows, were optimized to predict critical events from the communication data. A support vector machine then classified the data into categories with a high or low probability that a given time window includes a critical event. Using this approach, over 80% of the critical events were detected with an acceptably low false alarm rate. This model allowed the toolset to accurately detect critical events during a mission for inclusion in an AAR.

Team performance modeling was performed to predict the SME ratings of performance based on variables drawn from the text of the communications, such as semantic content, as well as variables drawn directly from the audio features of the communication, such as pitch, power and the Mel Frequency Cepstral Coefficient which were used to predict the presence of stress in speech. The best variables were selected to predict the team's performance on each of the five scales for each training event. The model's predictions were correlated with the SME ratings between .36 and .43, somewhat lower than the agreement between SMEs which ranged from .38 to .66 for single items.

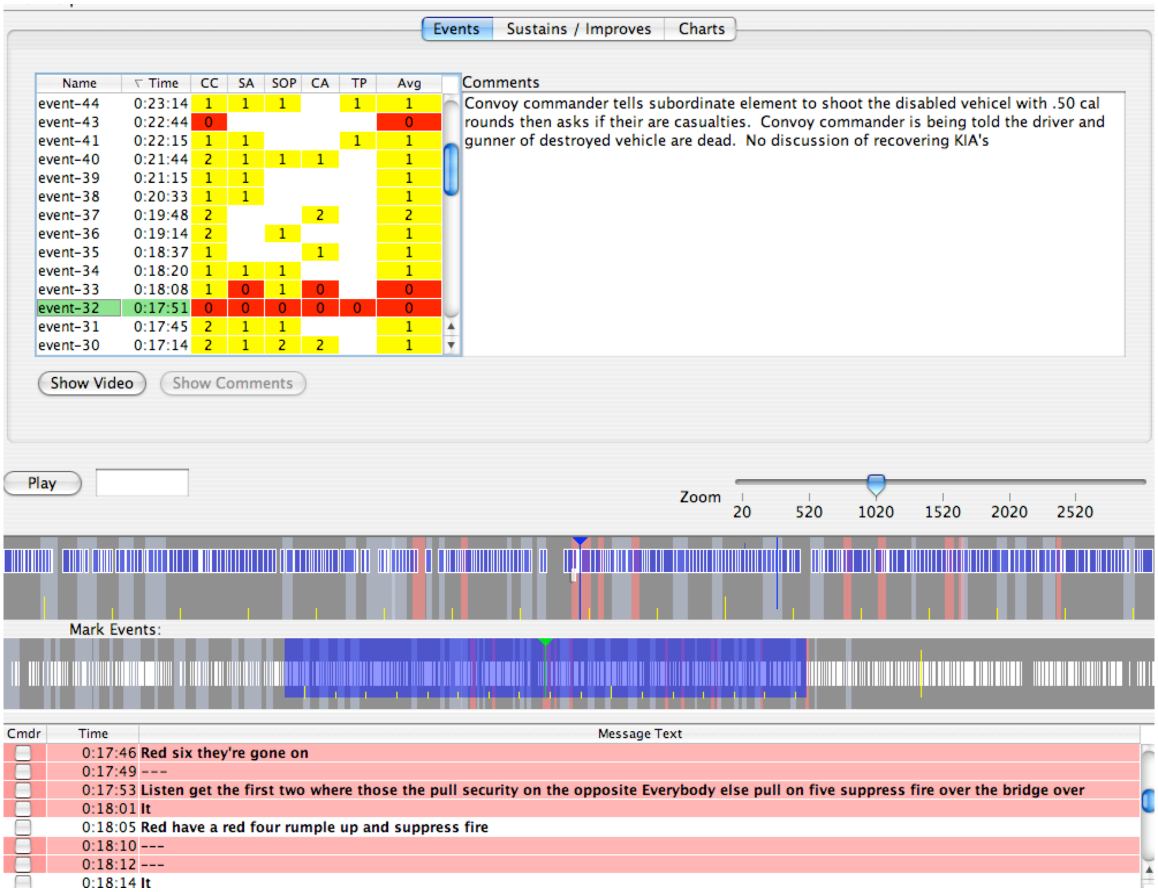
Team performance was also modeled for entire missions, instead of the separate training events in the missions, based on the ratings of the two SMEs with the highest agreement. Because the unit of analysis for this model was the entire mission, and the agreement results for the SMEs were reported using events as the level of analysis, additional agreement measures were calculated based on the team performance ratings for entire missions rated by both of the SMEs. The model's predictions correlated well with the SME ratings, with correlations ranging from .70 to .81 across the five scales, only slightly lower than the correlations between the two SMEs. Adjacent agreement between the SMEs and the model was also quite high, strongly supporting the use of the model in the toolset for assessing a team's performance.

## AAR TOOL DEVELOPMENT

Convoy training conducted at Fort Lewis using Ambush! and during STX lanes at NTC relies on the After Action Review process to maximize the benefits of training. During a well run AAR, the O/C or commander reviews the unit's performance, emphasizing areas where the unit would benefit from improvement as well as areas the unit should sustain at their current high level of performance.

The value of being able to provide a unit with recorded examples of their performance is unquestionable. After several hours of training, many team members may not be able to accurately recall a particular incident from earlier in training in sufficient detail to be able to learn well from their experiences. Currently, some video and audio from training events are collected at the NTC. However, the video and audio are seldom available to units for AARs or hot washes conducted in the field. NTC is in the process of installing the necessary infrastructure to provide live video and audio feeds to the O/Cs in the field, including laptops carried in the O/Cs' vehicles and plasma displays available in trailers distributed through the training area. These improvements will make it possible for O/Cs to use the recorded media of a unit's training to augment the AAR process. Within DARWARS Ambush! it is possible to record a unit's performance as they navigate the challenges in the virtual world, and then play the video back during an AAR. But two obstacles remain, even if all the multimedia is available. The first is the time required in finding events noted as training relevant during the mission by sifting through the video and audio recordings and making sure they cover the "teaching points" that illustrate a unit's weaknesses. With current O/C staffing shortages, the time that it takes to identify segments of video or audio of interest may overwhelm the benefits of using recorded performance for AARs. The second obstacle is that given the workload for understaffed O/Cs not all activity can be continuously monitored and critical events may be overlooked. By automatically analyzing the communications, this toolset extends the O/Cs reach.

The AAR tool we developed includes several functions to support O/Cs and commanders in preparing an AAR. As shown in Figure 2, O/Cs can view an entire training mission by events. This view provides a color-coded table of automatically selected events and critical events that are rated by the tool on the 5 scales: CC (Command and Control), SA (Situation Awareness), SOP (Standard Operating Procedures), CA (Critical Action Drills), and TP (overall Team Performance). The lowest scores are indicated by red, with the best scores shown in green, to help O/Cs spot events of interest. Clicking on the rating scale name (e.g. CC) sorts the events so the events with the best or worst performance on that scale will be visible at the top (see Figure 2), making it easy for an O/C to identify potential sustains and improves. Each event is linked to the audio recording, so clicking the event will play the associated audio files automatically, and show the ASRed transcript of the audio. Clicking the event will also display brief, automatically derived comments for each event that explain the event and ratings (see above right in Figure 2). As shown in the lower half of Figure 2, the display also allows O/Cs to browse using a timeline interface, with the ability to get an overview of the whole mission and zoom in to locate audio from particular parts of the mission they want to listen to.



| Name | ▿ Time | CC | SA | SOP | CA | TP | Avg |
|---|---|---|---|---|---|---|---|
| event-44 | 0:23:14 | 1 | 1 | 1 |  | 1 | 1 |
| event-43 | 0:22:44 | 0 |  |  |  |  | 0 |
| event-41 | 0:22:15 | 1 | 1 |  |  | 1 | 1 |
| event-40 | 0:21:44 | 2 | 1 | 1 | 1 |  | 1 |
| event-39 | 0:21:15 | 1 | 1 |  |  |  | 1 |
| event-38 | 0:20:33 | 1 | 1 |  |  |  | 1 |
| event-37 | 0:19:48 | 2 |  |  | 2 |  | 2 |
| event-36 | 0:19:14 | 2 |  | 1 |  |  | 1 |
| event-35 | 0:18:37 | 1 |  |  | 1 |  | 1 |
| event-34 | 0:18:20 | 1 | 1 | 1 |  |  | 1 |
| event-33 | 0:18:08 | 1 | 0 | 1 | 0 |  | 0 |
| event-32 | 0:17:51 | 0 | 0 | 0 | 0 | 0 | 0 |
| event-31 | 0:17:45 | 2 | 1 | 1 |  |  | 1 |
| event-30 | 0:17:14 | 2 | 1 | 2 | 2 |  | 1 |

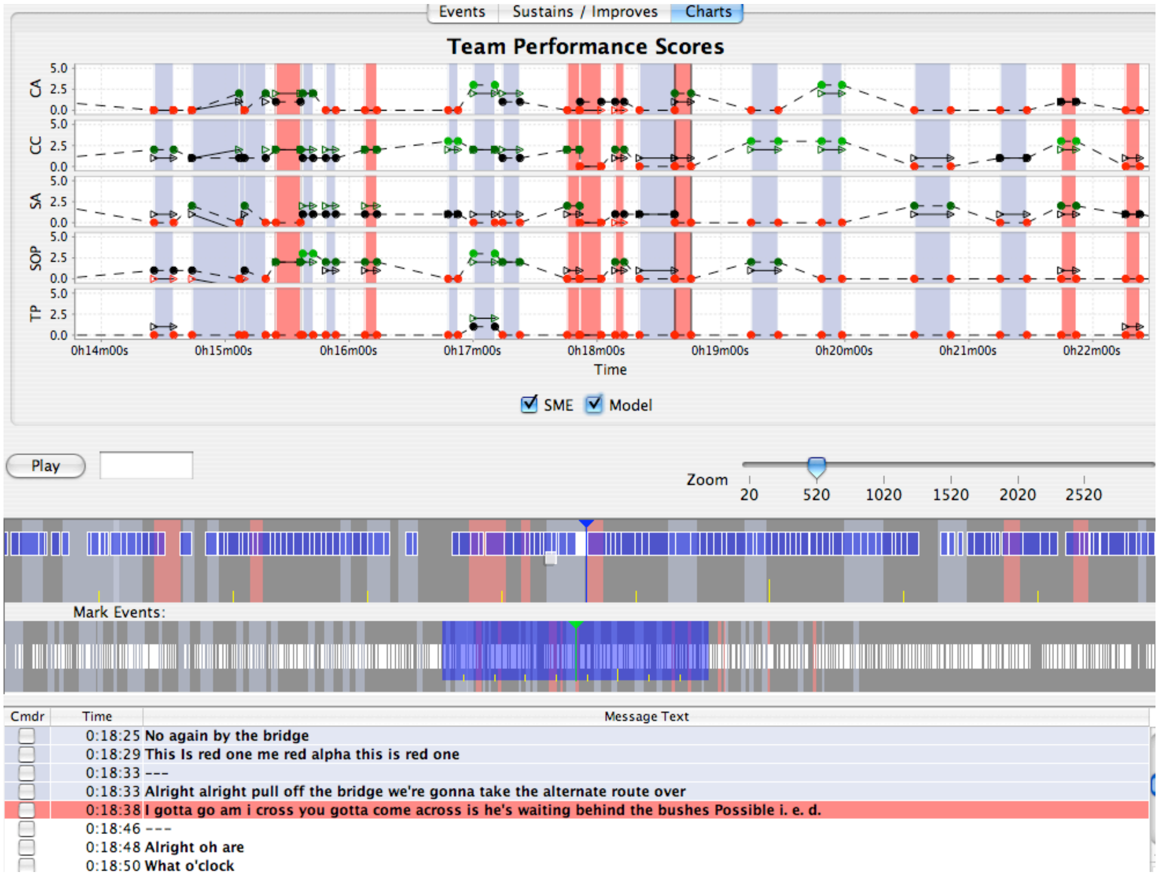
Comments

Convoy commander tells subordinate element to shoot the disabled vehicel with .50 cal rounds then asks if their are casualties. Convoy commander is being told the driver and gunner of destroyed vehicle are dead. No discussion of recovering KIA's

| Cmdr | Time | Message Text |
|---|---|---|
| □ | 0:17:46 | Red six they're gone on |
| □ | 0:17:49 | --- |
| □ | 0:17:53 | Listen get the first two where those the pull security on the opposite Everybody else pull on five suppress fire over the bridge over |
| □ | 0:18:01 | It |
| □ | 0:18:05 | Red have a red four rumple up and suppress fire |
| □ | 0:18:10 | --- |
| □ | 0:18:12 | --- |
| □ | 0:18:14 | It |

**Figure 2. AAR Tool Interface Showing Events and Ratings in a Table.**

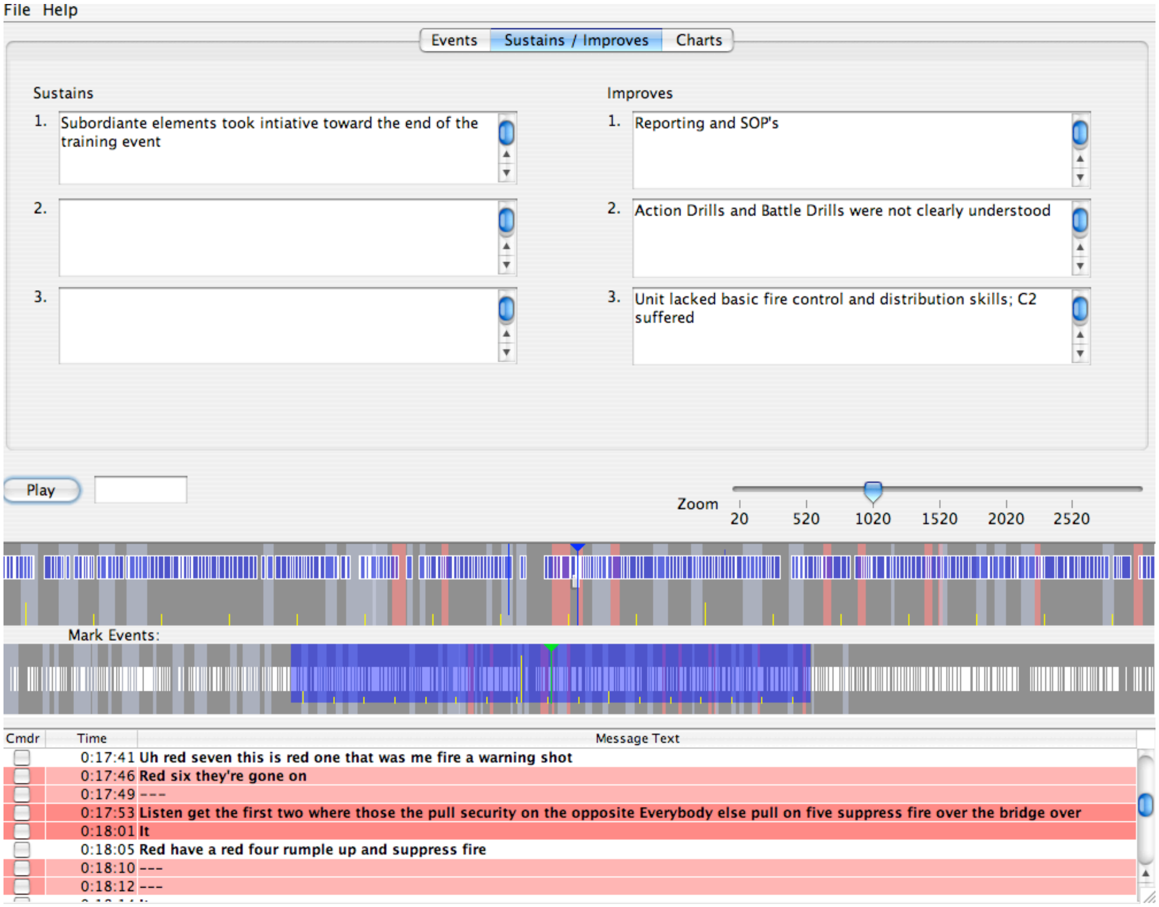
Figure 3 shows the chart view of a team's performance in the AAR tool. Continuous scores on each rating scale, generated by the tool, are displayed over the time course of the training mission. This display option allows O/Cs to view trends in performance, including improvements or declines in specific areas. For example, if a unit starts out having problems in command and control, but improves over the course of the mission, the O/C would be able to easily see that using this display.



| Cmdr | Time | Message Text |
|---|---|---|
| ☐ | 0:18:25 | No again by the bridge |
| ☐ | 0:18:29 | This Is red one me red alpha this is red one |
| ☐ | 0:18:33 | --- |
| ☐ | 0:18:33 | Alright alright pull off the bridge we're gonna take the alternate route over |
| ☐ | 0:18:38 | I gotta go am i cross you gotta come across is he's waiting behind the bushes Possible i. e. d. |
| ☐ | 0:18:46 | --- |
| ☐ | 0:18:48 | Alright oh are |
| ☐ | 0:18:50 | What o'clock |

**Figure 3. AAR Tool Interface Showing Events and Ratings with a Continuous Graph.**

As shown in Figure 4, the AAR tool also generates a set of suggested improves and sustains for each mission. These improves and sustains are automatically generated by the system, and are closely linked to the identification of events of interest. While O/Cs may prefer to select their own improves and sustains, this interface will provide possible options, and help prevent an O/C from missing an important training opportunity during an AAR.



Sustains

1. Subordiante elements took intiative toward the end of the training event
2.
3.

Improves

1. Reporting and SOP's
2. Action Drills and Battle Drills were not clearly understood
3. Unit lacked basic fire control and distribution skills; C2 suffered

| Cmdr | Time | Message Text |
|---|---|---|
| □ | 0:17:41 | Uh red seven this is red one that was me fire a warning shot |
| □ | 0:17:46 | Red six they're gone on |
| □ | 0:17:49 | --- |
| □ | 0:17:53 | Listen get the first two where those the pull security on the opposite Everybody else pull on five suppress fire over the bridge over |
| □ | 0:18:01 | It |
| □ | 0:18:05 | Red have a red four rumple up and suppress fire |
| □ | 0:18:10 | --- |
| □ | 0:18:12 | --- |

**Figure 4. AAR Tool Interface Showing Suggested Sustains and Improves**

### Evaluation of the AAR Tool

Two SMEs reviewed the AAR tool in order to provide us with feedback about its usefulness in supporting AARs, and to suggest improvements and other possible applications. The SMEs included our primary SME, LTC (Ret) Fena, and a second SME who had recently returned from his second tour as a convoy commander in Iraq. Both SMEs thought the AAR tool was valuable and would reduce the time required to prepare for an AAR as well as increase the scope of events that could be discussed. They emphasized that time is often the most precious commodity during training and the focus of the AAR tool should remain on shortening AAR prep time to maximize the tool's utility to O/Cs and commanders. Both SMEs thought that the tool layout was conducive to the way they would choose to use it to support an AAR. Specifically, they felt that the tool would allow a quick and easy three-step process for preparing an AAR:

1. Identify a unit's strengths and weakness at a glance, by scanning sorted event ratings;
2. Understand the weaknesses by examining these events in more detail, including listening to audio samples;
3. And last, pull all the information about the unit's performance together with their own comments.

The SMEs also had suggestions for improving the functionality of the AAR tool, including:

- Making critical events easier to find, either by creating a separate table for them or by marking them more clearly in the context of the other events;
- Allowing an O/C or commander to add their own brief comments to events and missions;
- Provide short descriptions of each event, such as "First IED" or "CASEVAC" to improve identification;
- Adding performance benchmarks to help standardize performance across units. They felt that rating a unit as "trained" on a particular metric, such as command and control, is often a subjective judgment, and the Army's training could benefit by

calibrating the ratings provided by the AAR Tool to a more objective standard.

The SMEs also believed that the tool could easily be extended to provide an O/C or commander support beyond a typical training mission AAR. Their ideas for extending the tool included adding longitudinal tracking to monitor a unit's performance over multiple missions. This would require archiving missions and adding tools to visualize and summarize performance over time. Benefits would include being able to identify performance trends, including recurring problems. The SMEs also felt that the tool could provide support for briefings up and down the chain of command, making it useful in a significantly wider variety of circumstances. Future work will include collecting additional feedback from representative users to insure that the continued development of the AAR is in line with O/C and commander needs.

## CONCLUSIONS

The feasibility of using this communication analysis approach was demonstrated for automatically detecting critical incidents, identifying performance changes, and evaluating team performance in both live and virtual training environments. Based on the success of this project, the AAR tool could be further developed into an operational tool for use in Ambush! and NTC STX lane training environments with some additional refinements.

The general approach used here translates well to other military applications requiring monitoring and assessment of teams. It allows near-real-time analysis and modeling of real (complex) communication data for networked teams. The combined toolset automatically models objective and subjective metrics of team performance and can generate its predictions within seconds. Because the models are automatically derived, the approach does not require large up front task analyses and instead capitalizes on the demonstrable strengths of O/C's AAR techniques. The toolset could be integrated into systems to monitor and provide feedback for teams, in both training and operational venues.

## ACKNOWLEDGEMENTS

We would like to acknowledge the contributions of the DARCAAT project team including Marita Franzke, Kyle Habermehl. Brent Halsey, Chuck Pannacione Manju Putcha, Jim Parker, Boulder Labs, David Diller, Laura Leets, Fred Flynn, Cyle Fena, Don Scott, Paul Asuncion, Reginald King, Brian Oman, Len Dannhaus, Nick Hatchel, Rick Travis, David Leyden, and Jamison Winchel. We would also like to thank the following organizations for participating in this work: DARPA DSO, ARI, Fort Lewis, NTC, and Fort Carson. This work was sponsored by the Defense Adanced Research Pojects Agency (DARPA) and the U.S. Army Research Institute (ARI).

## REFERENCES

Brannick, M.T. Salas, E. & Prince, C. (1997). Team performance assessment and measurement: Theory, methods, and applications. Mahwah, NJ: LEA.

Chu-Carroll, J. (1998). A Statistical Model for Discourse Act Recognition in Dialogue Interactions. Papers from the 1998 AAAI Spring Symposium. Jennifer Chu-Carroll and Nancy Green, Program Cochairs. 2001. Technical Report SS-98-01. Published by The AAAI Press, Menlo Park, California. Pp. 12-17.

Core, M. (1998). Analyzing and Predicting Patterns of DAMSL Utterance Tags. Papers from the 1998 AAAI Spring Symposium, Jennifer Chu-Carroll and Nancy Green, Program Cochairs, Technical Report SS-98-01, Published by The AAAI Press, Menlo Park, California. Pp. 18-24.

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. Journal of the American Society for Information Science, 41, 391-407.

Diller, D. E., Roberts, B., Blankenship, S. & Nielsen, D. (2004). DARWARS Ambush! – Authoring lessons learned in a training game. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference. Orlando, FL: I/ITSEC.

Diller, D. E., Roberts, B. & Willmuth, T. (2005). DARWARS Ambush! A case study in the adoption and evolution of a game-based convoy trainer with the U.S. Army. Presented at the Simulation Interoperability Standards Organization, 18-23 September.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. Behavior Research Methods, Instruments and Computers. 28(2), 197-202.

Foltz, P. W. (2005). Tools for Enhancing Team Performance through Automated Modeling of the Content of Team Discourse. In Proceedings of HCI International, 2005.

Foltz, P. W., Laham, R. D. & Derr, M. (2003). Automated Speech Recognition for Modeling Team Performance. In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.

Foltz, P. W., Martin, M. A., Abdelali, A., Rosenstein, M. B. & Oberbreckling, R. J. (2006). Automated

Team Discourse Modeling: Test of Performance and Generalization. In Proceedings of the 28th Annual Cognitive Science Conference.

Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. A. & Cooke, N. J. (2003). Evaluation of Latent Semantic Analysis-based measures of team communications content. In Proceedings of the 47th Annual Human Factors and Ergonomic Society Meeting.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J & Martin, M. J. (2002). Some promising results of communication-based automatic measures of team cognition. Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting.

Kiekel, P., Gorman, J., & Cooke, N. (2004). Measuring Speech Flow of Co-located and Distributed Command and Control Teams During a Communication Channel Glitch. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, 683-687.

Kuhn, C. (2004). National Training Center: Force-on-force convoy STX lane. Engineer: The professional bulletin for Army Engineers, April-June.

Laham, D., Bennett, W., & Derr, M. (2002). Latent Semantic Analysis for career field analysis and information operations. Paper presented at Interservice/Industry, Simulation and Education Conference (I/ITSEC), December 2-5, 2002. Orlando, FL.

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. Discourse Processes, 25(2&3), 259-284.

Landauer, T.K., Laham, D., Rehder, B., & Schreiner, M.E. (1997). How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans. In M.G. Shafto & P. Langley (Eds.), Proceedings of the 19th annual meeting of the Cognitive Science Society (pp. 412 417). Mawhwah, NJ: Erlbaum.

Landauer, T.K., Laham, D., & Foltz, P.W. (2000). The Intelligent Essay Assessor. IEEE Intelligent Systems 15(5) , 27-31.

Morrison, J. E. & Meliza, L. L. (1999). Foundations of the after action review process. (ARI Special Report 42), Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences, 82 pp.

Paris, C. R., Salas, E., Cannon- Bowers, J. A. (2001). Teamwork in multi-person systems: a review and analysis. Ergonomics, 43 (8), 1052-1075.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Van Ess-Dykema, C., & Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech, Computational Linguistics 26(3), 339-373.

U.S. Department of the Army. (2001, June). FM 3-0: Operations. Washington, DC.

U.S. Department of the Army. (2003, September). FM 7-1: Battle focused training. Washington, DC.