COLLANE: An experiment in computer-mediated tacit collaboration

Tomek Strzalkowski[1,2], Sarah Taylor[3], Samira Shaikh[1], Ben-Ami Lipetz[1], Hilda Hardy[1], Nick Webb[1], Tony Cresswell[4], Ting Liu[1], Min Wu[1], Yu Zhan[1], and Song Chen[1]

[1] ILS Institute, SUNY Albany, Albany, NY USA
[2] Institute of Computer Science, Polish Academy of Sciences, Poland
[3] Lockheed Martin Corporation, Arlington, VA USA
[4] Center for Technology in Government, SUNY Albany, NY USA

Abstract. We introduce COLLANE, an experimental collaborative analytic environment that allows a group of professional analysts to work together effectively on complex, multifaceted information problems. COLLANE has been developed to investigate innovative ways of harnessing the power of collaboration so that to maximize the quality of the analytical product while at the same time controlling for its hidden costs: bias, groupthink, compromise, suppression of dissent and individual initiative. The key innovation that we are advancing in this project is the concept of *ubiquitous tacit collaboration* enabled through computer-mediated information sharing between the participants. By design, tacit collaboration requires no extraneous effort from the users since the information exchange is both automatic and targeted to what each analyst is currently doing. It also requires no specific "engagement" with subject matter experts since their continuous virtual presence assures ubiquity of collaborative opportunities. In this paper we describe an initial prototype of COLLANE, explaining its basic functions and components.

Key words: collaborative analysis, information sharing

1 Introduction

Collaborative work can be both highly efficient and tremendously constraining. A dedicated team can often quickly solve a difficult problem that would stump an individual analyst for a long time. One key advantage of teamwork is its efficiency: a complex task can be subdivided among the participants into manageable subtasks that may be accomplished in parallel, matching individuals' strengths and capabilities. Another important advantage of a team is its diversity of ideas and viewpoints: in an optimal situation, the strongest, most plausible solution is created that reflects the contributions of all group members. However, those apparent strengths of collaboration are also sources of significant problems. For a team to deliver efficiency, the task must be divided into discrete, coherent pieces that align well with the capabilities of individual analysts, and achieving this requires skillful leadership. At the same time, too rigid a management

structure may easily drive the group to an early consensus by promoting groupthink and suppressing alternative ideas or less likely hypotheses. This is clearly an undesirable side effect, which is often considered crippling in investigative analytic tasks where plausible conclusions need to be drawn from fragmentary evidence.

It appears then, that collaborative work may be a mixed blessing unless new ways are found to take an advantage of it while avoiding the pitfalls. In looking for a suitable model it is instructive to observe how analysts, in the government as well as in business, law, and other investigative professions, organize their work. Until recently, these organizations have been traditionally depending upon the work of individual analysts who have deep knowledge and expertise in specific areas (countries, organizations, technologies, etc.) and who tend to work independently producing reports and analyses as tasked by their agencies. Of course analysts do not work alone; in fact, they interact constantly with one another seeking advice, bouncing off ideas, or looking for leads. Specifically, they often consult experts in areas where they may have less experience. From a traditional viewpoint, none of this normally counts as collaboration, since the analysts may have independent tasks for which they are individually responsible (and also receive individual credits). Nonetheless, this informal networking is a vital part of the information gathering process; it also has some hallmarks of "good" though indirect collaboration: pulling in multiple perspectives, including alternative views, and adding critical feedback, while also keeping the overall case management coherent and motivated. Can this model be replicated and expanded into a true collaboration? Can a new generation of analytic tools be designed through which tacit collaboration be harnessed and managed in a way that improves the quality of intelligence overall?

The COLLANE project has been established to address these problems and to develop a computer-assisted analytic environment that can support effective collaboration while avoiding the drawbacks associated with more traditional forms of teamwork. The model that we advance in COLLANE is *ubiquitous tacit collaboration* where we attempt to capture some of the benefits of informal networking mentioned above but without the disruption of having to stop one's work and call another person for advice. In our model tacit collaboration is more than networking though; it is true collaboration, focused on the task at hand to which all participants contribute, albeit indirectly. Tacit collaboration does not require the participants to subdivide their work or to actively coordinate their activities; instead, they are assumed to pursue their individual lines of analysis on the same or related problems. Collaboration occurs, tacitly, because the system: (a) captures the associative knowledge generated by each participant when they query data sources and retrieve and retain information; (b) keeps track of what each participant is doing at any given time; and (c) shares relevant information and knowledge among the participants based on its relevance, timeliness, and usefulness. This continuous targeted information sharing has an effect similar to having several colleagues walk into your office as if on a cue and offer the information and advice that you require at this precise moment; however, this

effect is achieved without the distraction normally associated with such activities. In other words, the relevant information is exchanged but no extraneous effort is needed to obtain it. As a result, the participants are aware of others' relevant activities and progress, past and present, which in turn informs and influences their own activities. Our hypothesis is that tacit collaboration, broadly defined, is more efficient and produces better quality analytic results than what can be achieved through individual work or through work in traditional open collaboration teams.

The COLLANE system has been developed to instantiate the above concept and to provide an experimental vehicle for exploring this and other forms of computer-assisted collaboration. The current, preliminary prototype can support up to 4 analysts working simultaneously on the same topic, and it incorporates basic information sharing capabilities sufficient for conducting meaningful evaluation experiments. To put things in perspective, the fully developed COLLANE system will eventually support a community of users and user groups working asynchronously on related topics. Furthermore, it will enable the automatic exchange of complex episodic and associative knowledge that is created by the participants' research activities. The current prototype was designed to support both tacit and open collaboration, as well as individual work by single analysts; this was essential for comparison between different work modes and also for deciding which of the system features need to be retained or expanded, and which new capabilities may be needed.

We have also designed an initial set of metrics for comparing both the efficiency and the effectiveness of each work mode, as well as for quantifying the user experience in each case. Some of the metrics were adopted from earlier evaluations conducted with single-user interactive information systems ([6]; [21]; [13]) to the extent that these metrics could be applicable in the collaborative setting. Nonetheless, developing a meaningful evaluation strategy for a collaborative system turned out to be a significant challenge. The focus of this paper is therefore as much on a description of COLLANE and the analytic experiments we conducted with it, as it is on the design of evaluation metrics that can effectively measure system performance in future experiments. Furthermore, due to a relatively small scale of the evaluation conducted to date, the results reported here can only be regarded as indicative of certain phenomena occurring in collaborative work. These will serve as a basis for developing more formal evaluations in the future.

In order to design a realistic exercise we turned to professional analysts representing various government agencies; we also asked these agencies to develop realistic analytic tasks. The analysts were presented with brief descriptions of problems, and asked to prepare comprehensive reports within a preset time limit. Analysts were using the COLLANE prototype through which they could search a fixed subset of web-mined data and collaborate.

The preliminary results from this study suggest that COLLANE-supported tacit collaboration has the potential to produce significantly higher quality analytic reports than would be possible when working alone or in open collaboration

groups. This assessment is not easy to quantify using the existing methods for measuring the quality of an information product, as we will elaborate further in this paper arguing for new quality metrics. A better information product does not simply mean finding the most relevant evidence (although it matters, of course), or even arriving at the most likely explanation of this evidence (while this definitely counts too). It also means alternative interpretations of what is relevant and how the different pieces interconnect, and moreover how these different interpretations stack and rank against one another. This latter effect is almost impossible to obtain in a single analyst environment, and it is very difficult to see in a traditional, open collaboration because it is normally driven towards a consensus.

The following is a summary of key observations from the analytic workshop with COLLANE. We elaborate each point in the rest of this paper.

- Information sharing improves the analytic process. Analysts working in collaborative teams (open and tacit) are exposed to more topic-relevant information than analysts working alone. Therefore, we may infer that the collaborating analysts are using more evidence to arrive at more informed conclusions.
- The quality of the reported intelligence improves. Qualitative assessment of the reports prepared by the collaborating teams suggests that better and more conclusions are drawn by collaborating analysts than by analysts working alone.
- Tacit collaboration improves the analytic process by introducing constructive competition. Analysts working in tacit collaboration tend to pursue alternative or complementary interpretations of evidence. Unlike in the traditional teamwork, they are not being driven into consensus or compromise.
- Tacit collaboration exposes multiple interpretations of the available evidence: the outcome is a diverse portfolio of reports that facilitates the survival of all sound hypotheses.
- Tacit collaboration improves analytic productivity by inducing analysts to do more useful work per time unit. Tacit collaboration allows the less experienced analysts to benefit from the more experienced ones without necessarily slowing them.

The study also revealed that current methods for measuring quality of information products, based primarily on content precision and coverage, need to be revised:

- Current evaluation methodology is inadequate because it is geared towards a single output of the analytic process (report quality), does not support the evaluation of multiple outcomes of tacit collaboration, and penalizes minority dissenting views.
- New evaluation metrics are required to assess the relative value of a multifaceted portfolio of reports and hypotheses covering a complex information problem. Ways are needed to rank the hypotheses and to quantify the information value of the set.

– Revised evaluation design is required in order to control for confounding factors such as level of experience, subject matter expertise, and analytic skills of the participants. In addition, we need ways of measuring the effects of differently skilled participants on team performance.

> ***No one really works alone. [Analysts] always work in a collaborative mode. [I need] to be able to ask collaborators, midway, take a look at my work and see if I am going in the right direction.*** (Analyst D, COLLANE Experiment, 2007)

2 The COLLANE System

2.1 Overview

COLLANE is a collaborative analytical environment designed to enable ubiquitous tacit collaboration among a group of analysts working on the same or related information problems. The current prototype also provides a platform for evaluating analytic effectiveness and for experimenting with various collaborative settings. The two key capabilities of COLLANE that enable effective collaborative work are: interactivity and information sharing. Interactive features include question answering, question refinement, answer negotiation, and data navigation capabilities, which can be accomplished through natural language dialogue as well as through a visual interface. Information sharing includes the creation and maintenance of a combined answer space, targeted and time sensitive delivery of relevant data items, as well as distillation and exchange of exploratory knowledge accumulated throughout the analytic session. This exploratory knowledge arises from analysts' information access, retrieval, and assessment activities that interlink queries, data items, and any relevance or utility tags assigned by the analysts to the data items they view. Subsequent *automatic* interchange of knowledge thus captured allows analysts to continuously take advantage of each other's *relevant* actions and insights. One must note in this context that relevant information and knowledge are not limited to what may be considered supporting evidence for a particular query but includes complementary, tangential, even contradictory items that may be part of alternative hypotheses advanced by other analysts.

Analysts using COLLANE may do so remotely and in an asynchronous manner. This extends the notion of tacit collaboration to situations where some of the participants may be offline or otherwise unavailable; nonetheless, the collected information and associative knowledge they leave behind remains accessible and sharable under the same rules as before. Furthermore, we can include in our design "collaboration" with legacy analyses completed in the past by people who are no longer around, as well as with an analyst's own prior work or alternative approaches. In order to manage the totality of information and knowledge created by such a complex collaborative effort, COLLANE maintains the Combined Answer Space (CAS), an efficient data storage, which is continuously updated and always accessible. Any analytical action initiated by one participant, such

as information search or relevance assessment, is automatically checked against CAS for the presence of any relevant items and their usage by other participants. This way the cognitive power of each analyst is maximized without creating an undue distraction or information overload.

2.2 COLLANE Design

COLLANE expands the interactive question answering technology in HITIQA ([15]; [12]) into a multi-channel, mixed-initiative interactivity that covers the entire analytical history of an information task. Unlike the more standard one-question/one-answer mode typical for the internet search engines such as Google, COLLANE can accept a series of interlocking questions keeping track of what the users have seen thus far. In addition to answering direct questions from the analysts, COLLANE acts as a coordinator and a facilitator, using various modes of interaction (both verbal and visual) to communicate similarities and differences among the information requested, collected, retained, and discarded by different analysts. Depending upon the session progress and the state of the emerging solution, the system may use different techniques, ranging from subtle to strong, to alert analysts of relevant activities by other participants in their group.

Any actions taken by an analyst while logged into COLLANE will effect changes in the Combined Answer Space. For example, new data items retrieved in response to the analyst's current questions are placed into CAS and their relevance to any previously logged questions, whether they came from this or another analyst are automatically assessed. The multi-modal dialogue manager (MDM) then decides how to notify each analyst about the changes that affect their individual workspaces. Some changes may be reflected in a dynamically evolving visualization, which can be immediately seen by all collaborating analysts, usually as minor background alterations. For other, more consequential changes (e.g., new or contradictory evidence) verbal alerts are used, i.e., a dialogue act is generated by the system in the form of a textual message, usually as a question or an offer that necessitates a response. COLLANE maintains a virtual individual working space for each analyst. In essence, an individual working space is a view of CAS in which the data items of interest to one analyst are made salient while all other data items are in the background or hidden from direct view. This allows MDM to conduct focused and meaningful interaction with each analyst rather than simply addressing them as a group. This interaction extends to the visual panels: each visual panel on a COLLANE client interface is a private view into the public space, reflecting a single user perspective. It allows an analyst to concentrate on his or her own work while also taking advantage of relevant aspects of other analysts' work. In particular, analysts may view and assess relevance of data items that belong to another analyst's primary view, thus altering their salience and indirectly affecting the other's workspace. It is the key function of MDM to make sure that such tacit collaboration has a positive effect on the performance of each analyst.

The analyst's interaction with the system occurs through a multi-modal dialogue that combines verbal (textual) exchanges and direct manipulations of the visual panel. As they work on a case, analysts ask questions, view system responses, save some items while ignoring others, inserting comments and annotations. By asking specific questions and by making certain decisions regarding the viewed data items, the analysts add their own knowledge, preferences, and biases to the system, and this knowledge, preferences, and biases are then shared with other analysts connected to the system. Such multi-way interaction has two effects: (1) it causes the individual workspaces to be populated with information related to their owners' queries, and (2) it alters the content of the Combined Answer Space. Unlike the localized changes in the individual work-spaces, any changes affecting the Combined Answer Space will be propagated to all participating analysts based on their relevance, thus further affecting their workspaces. Moreover, any explicit changes to one's individual view will immediately affect other views, although these effects may be only marginal, for example, other analysts may only notice changes that affect items currently visible to them. Depending upon the significance of these changes, the Dialogue Manager may utilize amplification messages to induce a desired reaction. The objective here is to keep all analysts current to the present state of interactions by facilitating but not forcing them to see each other's work and the effect of that work on the data.

The key advantages of the collaborative model outlined above are *completeness* and *efficiency*. Completeness is achieved through inclusion of multiple individual approaches and perspectives of a complex problem. Each analyst is now able to quickly identify alternative hypotheses to the present problem, by looking at the views over the same data created by other analysts' interactions. This helps, in turn, to identify where an analyst has perhaps missed certain evidence, or even where evidence directly contradicts the current working hypothesis. In this case, supporting documents can be identified, by looking into another analyst's folder. The Combined Answer Space plays a critical function in showing all emerging approaches in relation to one another. For example, the visual map of CAS may induce analysts to ask new questions, investigate new lines of inquiry or drop existing lines as no longer promising.

Efficiency is gained by accelerated collection and vetting of evidence. As all participating analysts work toward the same goal and share their insights and partial results, we gain the effect of parallel processing so that much of the duplication of work common in individual work settings can be avoided. Furthermore, since COLLANE does not preselect any specific roles for the analysts, the team can maintain dynamic flexibility by allowing each participant to take the initiative and explore avenues they find most promising. By asking questions to follow up competing or complementary hypotheses, the analysts effect changes on the combined answer space, thus *rapidly* (but indirectly) communicating the effects of their actions to others.

This model also significantly increases the *quality* of analytical products (which is subject to experimental verification, discussed in a later section).

One reason for this is the expanded evidential basis for the team report where each piece of evidence has been thoroughly vetted as potentially competing hypotheses are considered. This can be contrasted with one-analyst/one-system setting where the analyst advances a hypothesis, and the system then retrieves supporting or refuting evidence but it cannot on its own come up with an alternative. Another reason for increased quality may be the combined expertise brought into the task by all participating analysts. Experienced senior analysts are expected to exert considerable guiding influence on junior team members thus improving the quality of their work without overt instruction. A further reason for increased quality is the competitive element that promotes diversity of approaches in tacit collaboration: analysts are not driven towards a consensus or a compromise; instead, a portfolio of alternative and complementary analyses is generated.

2.3 Key Functionalities

The overall architecture of COLLANE is illustrated in Fig. 1. CAS is the central shared working space on any given information task; however, it is actively managed by the Dialogue Manager (MDM), which maintains individual views to best support each analyst. For clarity, we omit some components that have not been integrated yet, but will be discussed briefly later.

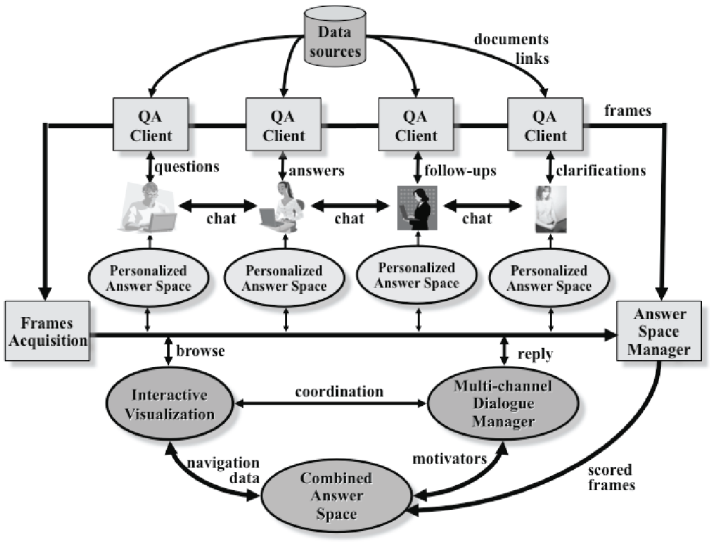

Fig. 1. COLLANE-1 system architecture

Combined Answer Space. The Combined Answer Space contains all information units under consideration by a group of analysts working on an information task (a case). Each time an information request is initiated by any of the analysts, all data items matching this request will be pulled out of available data sources. These data items, currently text snippets but in the future also multimedia files, are normally expected to have different degrees of relevance to the analytical problem at hand, as well as to the questions that analysts pose through their client interfaces. The initial assessment of relevance is computed by the system in response to specific questions, but it may be subsequently modified through interaction and other analyst actions such as saving an item in a "shoebox" or a draft report. Due to differences of opinions and approaches between the analysts, this scoring system is inherently multi-valued.

CAS is built out of all the retrieved information units, not just those that may be used by analysts in their final reports. In order to facilitate uniform handling of all information types by the system, each information unit (or a group of like units) is assigned one or more *event frames*. Event frames are template-like structures that represent the content of the underlying information unit: usually an event (e.g., an agreement, a transfer, an attack, etc.). Frames are classified into a dozen or so types, which are automatically defined for each subject domain (e.g., trade, politics, terrorism, emerging technologies, etc).

Each frame provides an "access handle" to the original information unit through which the system can compare and manipulate information content *regardless of their origins*. Our experience has thus far been primarily with text-extracted frames; however, similar structures can be extracted from e.g., video clips, after which they can be handled transparently by COLLANE.

A frame represents only a portion of the information contained in the original unit: the predicate, key attributes and selected modal operators. For example, *ATTACK(X,Y,Z)* represents an event where X attacks Y using (weapon) Z. Additional attributes specifying time, location, and modality (e.g., past, future, alleged, denied, etc.) are extracted as well; *for instance, In northern Baghdad, the owner of an ice cream shop was shot dead outside his store on Sunday morning...* We have developed an ontology of basic event frames that cover a number of analytic domains. For example, the weapons proliferation domain includes several basic events that characterize this domain from a national security viewpoint: *TRANSFER, DEVELOP, AGREE,* and *ATTACK*. These basic event frames instantiate to specific events reported in the information sources; for example, Iraq importing uranium from France would be an instance of *TRANSFER* frame, as shown in Fig. 2. Other domains will use similar sets of basic relations.[5]

CAS is built and managed by COLLANE as the case analysis progresses, and over time, it can become quite complex. Below is a summary of the key elements:

[5] For more details about the system of event frames and how they are acquired, the reader is referred to ([5]).

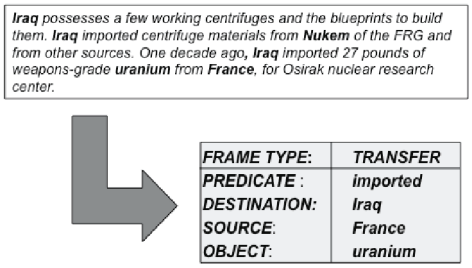

Fig. 2. A text snippet is assigned a TRANSFER frame

1. *Frames representing the retrieved data*: (text passages, XML snippets, other media types) with one or more frames per information unit. In addition, frames representing the same exact event are merged with their attributes combined, in which case a single frame may represent a group of information units.
2. *Relevance scores for each item*: Since information units may be retrieved in response to different questions, multiple scores are assigned to each along with the corresponding pedigree (which question, whose question, any amendments). Relevance is then assessed to all questions posed, including those answered previously. In addition, any direct actions by any analyst with respect to a particular item (mark as relevant, mark as non-relevant) are captured. This allows COLLANE to calculate and display a combined score of each item with respect to the overall analytic case.
3. *"Ownership" information*: Items retrieved in response to questions posed by individual analysts are assigned to them so that an individual answer space can be identified. Clearly, such individual spaces may overlap in various ways, but it is important for an analyst to know where their work is located vs. other analysts. Furthermore, information sharing is expected to be more effective when it is passed along with such essential context as "who's got it?", "who's seen it?", "what they did with it?", etc.
4. *Cross frame links*: While this feature is not currently implemented, frames will be linked by shared attributes forming various chains: temporal, geospatial, person/organization. Additional linkages may be inserted by link analysis components external to COLLANE (e.g., social, communication, etc.)

A schematic (and highly simplified) illustration of CAS is shown in Fig. 3. We should note that the overall information model does not have to be consistent—inconsistencies are expected to arise from analysts pursuing different approaches and forming incompatible hypotheses. In particular, relevance assessments for each data item will likely vary between analysts for a variety of reasons, including differences of opinion but also utility of a particular item to each analyst's workspace.

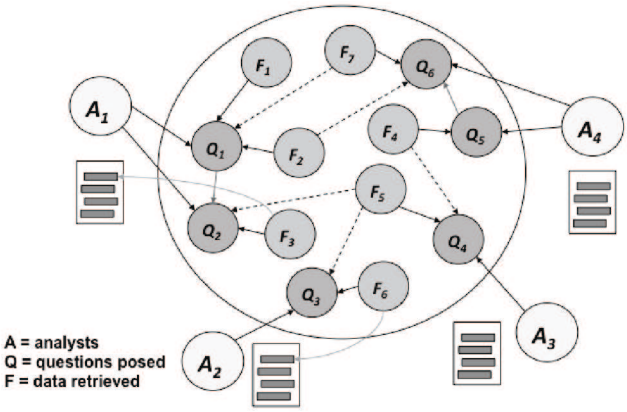

Fig. 3. A schematic illustration of CAS showing questions posed and data items collected by 4 analysts. Links between question nodes (Qn) indicate order in which they were asked; links between frames (Fn) representing data items and query nodes indicate relevance.

Multiple Views of CAS. The Combined Answer Space holds the entire evidential history of an analytic case. While all information is available and persistent, it is not necessarily viewable all at once. Instead, *multiple views* of it are created as required to support each analyst. In a typical case, each analyst's primary workspace is in focus while the remaining parts of the Combined Answer Space form a background. This is necessary to communicate its content effectively, whether by graphical means (visualization techniques) or by verbal dialogue. Here is how COLLANE uses the Combined Answer Space to interact and to facilitate teamwork:

1. *Supporting Dialogue Manager*: CAS is the primary data structure supporting the Multi-channel Dialogue Manager. Much of the dialogue generated by the system arises in order to resolve any perceived inconsistencies in the model as well as to negotiate the scope of the answer space.
2. *Rendering into interactive visual display*: For each analyst, her/his workspace is displayed and organized around the questions they ask. The display includes only these data elements that are necessary to answer the questions and to support effective interaction.
3. *Rotating views*: An analyst may "rotate" available views to examine other analysts' work-spaces. This feature is only partially implemented at this time. The plan is to support switching from one analyst's scenario-level view to another's. While in another workspace, the analyst's own data items can be viewed in relation to the items collected by the other analyst.
4. *Browsing*: Analysts can browse the visual display, access original data items, change/add their relevance assessment, and copy them into their reports (or

"shoeboxes"). Changing relevance assessment of a data item provides a direct feedback from a user, and this information is propagated to other users depending upon their individual circumstances and using an appropriate communication channel (as discussed further below).

5. *Analyzing information from retrieved evidence, chat messages and user copies*: In addition to matching retrieved text passages (or other media types) with user questions, CAS keeps track of questions and statements transmitted through COLLANE's chat interface, which allows analysts to communicate directly if they so choose. If an analyst's chat statement, for example, is found to match a question some other analyst has posed to the system, CAS and the MDM will share that piece of information with the other analyst. Likewise, CAS will search among retrieved items and present to the user any that are relevant to questions he or she has posed through chat. If an analyst finds and copies a passage of text from a full-document link (that is, a passage not identified by the system as highly relevant), CAS will analyze the new passage and look for matches among questions already asked, thereby enriching the answer space for other analysts.

Hypothesis Footprints. A natural consequence of the collaborative design outlined in the previous sections is the multi-dimensionality of the analytic process in COLLANE. Each analyst on the team may pursue a different strategy and consider alternative, even contradictory hypotheses. These hypotheses are not necessarily apparent to an observer, as the analysts may be exploring various options that appear promising at one time or another. The totality of the analyst's actions while pursuing a hypothesis: evidence collected, questions asked to collect it, assessment of this evidence for relevance, responses to system suggestions, and reactions to other analysts' progress—all these elements form an information "footprint" left by this analyst while pursuing the hypothesis. In COLLANE we define a concept of *hypothesis footprint* to be the set of all actions performed by an individual analyst while considering a specific hypothesis.

The main advantage of computing hypothesis footprints is the ability to recognize that analysts may be pursuing different approaches to a problem. We should note that this is only meaningful in a collaborative environment, where analysts can be made immediately aware of such alternative approaches. The system is unlikely to guess, based solely on the information in the footprint, which hypothesis is being considered, but radically different footprints left by two analysts may signal that they are pursuing different approaches. The system may now attempt to reconcile these differences by making the analysts aware of each other's progress, thus further accelerating the analysis.

In the course of their work on a case, each analyst may pursue several hypotheses, thus an additional technical challenge is to determine where one hypothesis ends and another begins. Moreover, while a particular footprint cannot be used to prove or disprove a hypothesis, we might be able to discern from the analyst's actions (saves vs. discards, line of questions, etc.) whether he or she succeeded or abandoned an approach. This aspect of COLLANE is not

yet fully implemented. In its current form a hypothesis footprint is represented as an undirected acyclic graph (Fig. 4) with analyst's actions as nodes and the data items associated with those actions as leaves. This structure allows for swift and straightforward comparisons between two distinct footprints, as well as for finding characteristic episodes within a single footprint that may indicate approach boundaries.

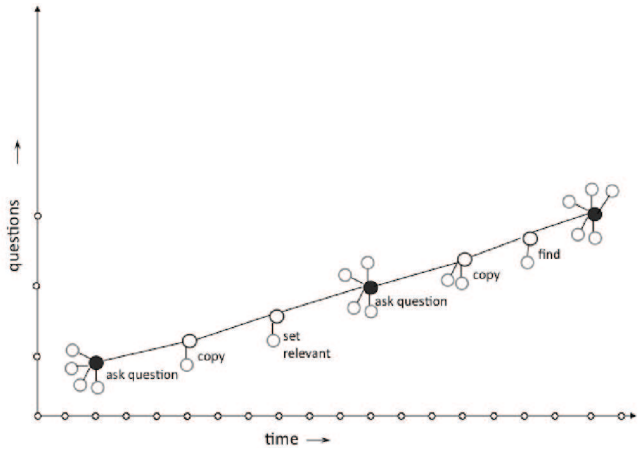

Fig. 4. Graphical representation of an analyst's Hypothesis Footprint

Multi-channel Interactivity. A key aspect of COLLANE is its interactivity, which allows for efficient information exchange between the analyst and the system. Analysts may negotiate the exact scope of each question, receive suggestions on how a question may be reformulated or expanded, and be alerted about any related or contradictory information found. Interactivity facilitated by the system also encompasses direct and indirect communication between the participating analysts. Indirect communication occurs when one analyst's actions affect the workspace of another analyst, which may in turn cause the second analyst to alter his or her working hypothesis. The interaction may proceed verbally or visually with the Dialogue Manager selecting the optimal means depending upon the urgency of a communication, and other contextual parameters.

Given a complex information problem, an analyst would normally pose a series of questions, probing for specific details that could support an unstated hypothesis or else open avenues for further exploration. The choice of a particular line of questions may reflect the analyst's background, prior knowledge of the subject matter, or other biases. Each question may be compared to a narrow spotlight shined into massive data, which means that an answer, even if correct,

produces only a "dot" of information at a time. Therefore what questions are asked and how they are asked can make a huge difference in the final outcome.

One way to improve the odds of finding and connecting the right "dots" of information is by adding a broader context "halo" around each answer returned and this can be accomplished by engaging the analyst in a *dialogue* to evaluate additional information items that appear highly related to the direct answer. This has the effect of providing a broader evidence context to the answer reported and increasing the analyst's success rate. Even if the contextual information is not relevant to the problem at hand, its presence may strengthen the answer selected; i.e., it may indicate that further exploration in a particular direction is unlikely to be useful. We have accumulated experimental evidence based on our work with active duty analysts that this contextual dialogue increases both the speed and the quality of the analysis and frequently leads to additional information nuggets that analysts utilize in their reports ([15]; [20]; [21]; [9]; [6]).

This human-machine interactive analysis is significantly accelerated in a *multi-channel, multi-thread dialogue* situation. When multiple analysts approach the same problem, they are likely to do so from different perspectives, thus cutting different evidence paths though the massive data. Each of these evidence paths may reflect the pursuit of an alternative hypothesis, thus becoming a hypothesis footprint, as discussed above. By creating a particular hypothesis footprint, an analyst communicates to COLLANE, and indirectly to other participating analysts, that they are pursuing a particular approach. When these footprints are combined and compared, a significantly larger, multi-dimensional evidence base is obtained. This provides a much wider context for each data item that the system may now provide to each analyst, creating even more opportunities to continue the search. Still more importantly, from the perspective of any one analyst on the team, the system is now significantly more responsive and *forthcoming*: it provides active feedback to all direct questions asked and also explicates alternative explanations for evidence, based on what other analysts are doing.

In COLLANE's collaborative environment, the quantity of information relevant to an individual analyst rapidly increases, and we require an efficient interaction mechanism without overwhelming the users with streams of communication. The key function of the Multi-Channel Dialogue Manager (MDM), in addition to "regular" human-computer interaction support, is to alert each user to new information as well as new, promising lines of investigation relevant to their enquiry, as they are being uncovered by other collaborating analysts.

Several interaction decisions are taken by MDM as to how, and when, to alert the user about developments outside their individual workspaces. For one thing, COLLANE will only engage in a dialogue when the nature of the information is such that intrusion into analysts' current work process seems warranted. For example, Analyst A has asked a question some time ago. Analyst B now asks a new question, which results in new data items, some of which match the original question of Analyst A. Our decision on how to handle this new, matching information depends on what Analyst A saw in response to

the original question. If the original answer appeared satisfactory (i.e., Analyst A saved new information into their "shoebox" or a report) then the update to A's working space will be silent and unobtrusive. For example, relevant items are silently dropped into appropriately labeled folders in the analyst's workspace, and a visual "flag" is raised over the folder, mailbox style, to indicate a new item arrival. On the other hand, if the original answer was not satisfactory, as evidenced by lack of copied material and possibly several fruitless followup questions, then a more visible "highlighting" of relevant folders in the visual workspace is used to alert the analyst that the data item he was unable to find has been located and may be viewed now. In an extreme case when a newly discovered data item appears to contradict some earlier findings, or is an entirely new data item, where no previous data was seen, the system may engage the user in a direct verbal dialogue.

(1)Analyst A: *Is there any evidence that man-made artificial reefs are beneficial?*
(2)COLLANE: Displays matching results to Analyst A
(3)Analyst B: *Where has sea life increased due to artificial reefs being constructed?*
(4)COLLANE: Displays matching results to Analyst B, including some results deemed relevant from the previous question of Analyst A
(5)COLLANE: Displays to Analyst A any new results retrieved by Analyst B

Fig. 5. Interaction between COLLANE and 2 analysts

In the example in Fig. 5, Analyst B asked a question which included more specific details (possibly based on prior and tacit knowledge) than the initial question of Analyst A. By using this specific hypothesis, sea life increase, COLLANE is able to retrieve new information for *both* analysts. How this new information is displayed to Analyst A depends on their prior actions. If no relevant data was seen in response to the initial question (1), COLLANE will initiate verbal dialogue at step (5), directly informing Analyst A of new, directly relevant information. If the new information complements a partial answer, COLLANE may choose a less disruptive notification through visual display, e.g., raising a "new arrivals" flag on a folder.

More generally, given an emerging solution to an analytic problem, the system employs a series of dialogue moves, which may be either verbal or visual, in order to draw the analyst's attention to a particular detail, or an issue, or some changes that may be occurring. A dialogue move is a particular manner of communicating information, which also aims at eliciting a desired reaction from the user. Some dialogue moves are more direct than others, e.g., a direct question usually compels the other party to respond ("*Would you be interested in information on new marine habitats?*"), while an open-ended offer may be

ignored or put aside ("*Please check these when you have a moment*"). The selection of which dialogue move to employ and when to employ it is all-important because the dialogue should never become a distraction or nuisance to the analysts. A combination of verbal and visual communications gives the system significantly more options that could also be deployed simultaneously. For example, a continuous complex change in data may be more readily visualized than described.

Timing is also important for such actions to achieve the desired effect. The analyst may have aborted an earlier line of questioning altogether, so we must be careful as to how we inform them, not to presume that the new information is still vital to the completion of the task. This is where the tracking of hypothesis footprints becomes critical—a feature that allows COLLANE to tell which questions remain open and which are no longer active. The Combined Answer Space provides the structure necessary for efficient dialogue with all users.

To summarize, the role of the MDM module is to accept direct queries from each user, decide what additional information is needed, when it is needed, and how to get it: by asking the user, by observing the user, or by inducing some action from the user. The system continually measures disconnection between the user's interpretation of the analytical problem (a hypothesis) and the content of the answer space obtained (the evidence). This can manifest itself as a mismatch between the questions posed, the relevant information found, and the information retained by the analyst. The objective is to make the user an active and effective participant in the information-seeking process, but to do so in a manner that is unobtrusive and naturally fits with the task flow.

The approach described above should be contrasted with more standard information system approaches to interaction. For example, most current "interactive" information systems implement only very basic forms of interactivity, typically variants of passive feedback. Document retrieval engines, including Google, are good examples of this: the user must decide if and how to revise the query to get better results. In dialogue research, early forms of interactive systems used fixed menus to guide users through a maze of options often unrelated to the user's information need. While theoretical research on dialogue modeling has made good progress ([1]; [18]; [8]; [22]; [23]), the practical implementation still lags behind. A significant development was the AMITIES project ([4]), which delivered a practical implementation of the data-driven dialogue approach, and was subsequently adapted to the unstructured data in HITIQA ([17]; [11]). Another related area is research on multi-party dialogue (e.g., [19]; [7]; [3]); however, this work concentrates primarily on the structural and functional aspects of the interaction, rather than on the information exchange, which is critical for COLLANE.

Direct Communication among Analysts. To facilitate inter-analyst communication, COLLANE provides a chat mechanism by which analysts (or teams of analysts) can communicate directly. Based on the experiments we conducted, analysts use chat primarily to bounce off ideas and to ask questions of each other

related to the current scenario, e.g., to inquire if others had a better luck with a particular topic or exchange prior knowledge about a topic, etc. Other uses we noted include various forms of work coordination among analysts. While COLLANE does not require analysts to collaborate openly, this direct communication channel allows them to subdivide a complex task and to exchange suggestions and advice on the progress thus far. The chat channel also complements tacit collaboration particularly when information sharing appears slow ("*Can't find anything on opposition to cabinet restructuring, did you?*").

COLLANE considers chat exchanges between analysts as another source of information that may reveal analysts' prior knowledge of a topic as well as other assumptions they make. Using discourse analysis tools, such as a general domain Dialogue Act tagging mechanism ([22]), COLLANE spots key excerpts in this information interchange; specifically we identify classes of questions and statements relating to known types of named entities, e.g., people, locations, organizations, etc. In a statement, we are looking for novel data items that may represent tacit knowledge exchanged between collaborating analysts: such knowledge is captured into the CAS, although it is also clearly identified as having originated in inter-analyst chat. For questions posed through chat, COLLANE will search available data sources for candidate answers, as well as the CAS for similar questions already answered by other analysts. In this way, COLLANE augments analyst collection ability by complementing the external data sources with the shared knowledge built by the collaborating team, currently and in the past.

Intuitive Conceptual Visualization. The role of visualization in COLLANE is twofold:

- The primary role is to create a representation of the Combined Answer Space that would allow the global view of all collected information and analyst individual views to coexist on an interactive display;
- The secondary role (but no less important) is to extend the capabilities of human-machine dialogue by allowing a greater variety of means of communication: non-linear dialogue acts, low-disturbance messages, and subtler alerts.

Our main objective in COLLANE has thus far been to develop a conceptual design for the visualization interface. In the future, we plan to develop a more advanced graphical rendering. Based on a series of user-centric experiments, we have identified the following requirements for the effective interactive visualization required to support COLLANE:

1. Effective visualization must clearly communicate the current content of the Combined Answer Space and the progress of the analysis.
2. The visualization must let analysts alternate between wide (more context) and narrow (individual workspace; specific aspect subspace) views of the answer space.

3. The visualization must allow for easy viewpoint "rotations" so that different analyst's views can be switched to as needed.
4. The visualizations must complement the verbal (text window) dialogue. Ideally, both means of communication should mix seamlessly and naturally.

In the current version of COLLANE, the default view of the combined answer space is associated with the most recent question that an analyst posed to the system. This view shows all information units considered at least partly relevant to the question including these just retrieved and others that may have been found previously by other analysts. It supports the analyst's current focus and also allows for detailed content negotiations to occur via dialogue. Figure 6 illustrates this; the reader should note that the icons (representing individual information units) are organized into groups (e.g., by event types) but not necessarily by relative semantic compatibility or "distance", which are difficult to determine objectively. The focus is thus on clarity and ease of perception.

In Figure 6, we note that a user question "*When was Teflon invented*?" returned a number of information units that fall into 6 groups: two groups of development events (DEV label), one group of transfer events (TRF label), a single attack event (ATT label) and two groups of other general events (GEN label). Color-coding represents the degree of relevance computed by the system (discussed further below), while the icon shape identifies the source (direct retrieval or shared from another analyst). The answer found by COLLANE's integrated QA system appears on the right panel. The analyst may change the view to any previous question by bringing it into focus from the answer folders panel (pictured in the screenshot on the right side of Figure 6).

From the question-level view (left side of Figure 6) the analyst may zoom in on each of the event groups thus entering an event (or frame) view (right side of Fig. 6). At this level additional details about each event frame are displayed, including key attributes, such as Agent, Object, Time, Location, etc. The analyst may also consult the underlying data sources and toggle the relevance assessment of each item.

As noted above, icons on the visual panel represent information items (text passages) that are delivered to the analyst in response to the most recent question. These items may come from any number of external sources that COLLANE may search as well as from the data collected by other collaborating analysts in the Combined Answer Space. We briefly explain the significance of colors and shapes of the icons. All icons represent salient events and relationships found in the source text passages, with key attributes displayed on the periphery. The color indicates the degree of relevance to the user question as estimated by COLLANE algorithms: the dark blue are most relevant, the red are least relevant. The dark blue items are automatically saved into answer folders, while the red items are considered not relevant although potentially useful for contrast or as context. The intermediate colors (light blue, green, yellow, orange, etc.) indicate items with increasing numbers of conflicts with the question, which nonetheless may prove relevant upon closer examination. As already discussed above, COLLANE may engage the analyst in an explicit

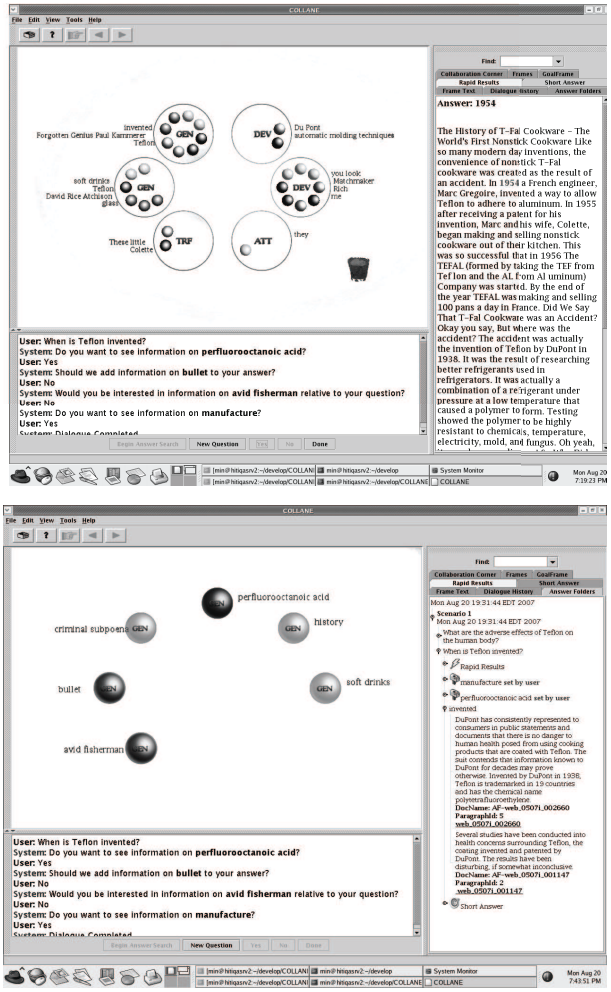

Fig. 6. Individual workspace views with frames grouped by event type and key attribute (top display) and individual data icons inside one of the groups (bottom display). Color-coding indicates the degree of relevance to the question (dark blue—seen as the darkest shaded icons—are most relevant)

dialogue in order to clarify the relevance of some of these items; however, the color system itself is a form of silent (visual) dialogue about items which are left for the analyst to examine.

In addition to colors, COLLANE uses icon shapes to represent the origins of the information. Circular icons (shown in the pictures in Fig. 6) indicate original new items directly retrieved from external data sources. Triangular icons (not

shown) represent related items found by other analysts either in the past, or concurrently, or possibly in the future (this latter possibility arises when the analyst reexamines the answer to a prior question). Currently, only the most relevant items from other analysts' workspaces are displayed, i.e., the items that would receive dark-blue coloring on the visual panel.

From the question level the analysts may also zoom out to the Scenario-level view where they can see all their questions posed in connection with the current task. This view provides a wider perspective on the work done up to this point and helps the analyst to assess the state of task completion. At this level it is also possible to meaningfully compare the progress with other analysts on the same team. This can be accomplished simply by switching to the scenario views of other analysts and noting the questions they posed, the answers they obtained, as well as their assessment of any common data items. The scenario view is currently under development and has not been included in the version of COLLANE we tested. It is likely to display multiple question groups in a reduced resolution that can be zoomed into by passing the mouse cursor over them.

3 Experimental Evaluation

The initial prototype of COLLANE described in the preceding sections was built during the first year of the CASE Program to support up to 4 analysts working simultaneously. While the research and implementation process has only begun, the system's development reached the point where direct feedback from potential users was required to assess the progress made thus far and to prioritize the challenges lying ahead. To do so, we have organized, in close collaboration with the U.S. National Institute of Standards and Technology and other government organizations, a collaborative analytical exercise to evaluate COLLANE performance on realistic information analytic tasks. The primary purpose of this exercise was to assess whether our testing methodology can support a meaningful evaluation of collaborative systems in general and COLLANE in particular. The secondary objective was to obtain a preliminary measurement of effectiveness of information sharing in COLLANE.

In September 2007, the team has conducted a five-day on-site analytical workshop with a group of 8 professional analysts representing various information services of the U.S. Government. During the workshop the analysts were presented with a series of realistic information problems of strategic nature and asked to prepare draft reports on each problem within a preset time limit. Analysts were divided into several groups and each group used COLLANE to collect and organize information, to prepare one or more draft reports, and, when appropriate, to collaborate. Each group worked under different conditions: open collaboration, tacit collaboration, or individually, as will be explained in more detail below. The searchable dataset consisted of approximately 2 GB of text documents premined from the Internet. We were interested in comparing the quality of reports produced by each group, as well as the effort expended and the user satisfaction; specifically:

1. Evaluating efficacy of tacit collaboration technology, specifically automated, targeted information sharing, in solving complex information tasks under limited time and resource conditions.
2. Comparing several collaborative and individual work settings and how they affect the use of the technology and the outcome of the analysis.
3. Determining if the current evaluation design is feasible and sufficient to measure the impact of various forms of collaboration on the analytic process and on the quality of the results.
4. Gaining insight into how the current COLLANE technology needs to be advanced to obtain a more effective tool.

3.1 Overall Evaluation Principles

Our objective was to design a methodology for evaluating the effectiveness of collaborative systems, such as COLLANE, for solving complex information problems by teams of analysts. This takes into account the following key dimensions:

1. *Quality of Solution*: an objective measure of how well the problem has been solved. This includes importance (criticality), coverage (completeness), precision (non-redundancy), and organization of the final report. This quality can be assessed by a panel of experts.
2. *Quality of the Process*: an objective measure of how the analytic process is affected by the technology. This includes accuracy of intermediate steps, rate and timeliness of information sharing, effectiveness of dialogue, etc. This also includes more complex measures such as the number and quality of hypotheses considered, the depth of the information search, and the rigor of the attempts to prove or disprove hypotheses, etc. This quality is measured by a combination of standard accuracy metrics (recall, precision, MRR) and the analysis of the structure of the interaction logs left by each analyst.
3. *Effort Expended*: an objective measure of the user effort expended to obtain the solution. This includes elapsed time, the number and types of steps required, number of sources consulted, etc. Effort is estimated from system logs that capture all significant task events and time stamps.
4. *User Satisfaction*: A subjective perception of difficulty of the process and confidence in the resulting solution. User satisfaction can be measured along many dimensions primarily through specially designed questionnaires.

We note that the above dimensions are partially orthogonal and which of them is more important depends upon the nature of the task. In most tasks the report quality will likely dominate other criteria; however, the process quality may be a necessary pre-requisite, i.e., it is hard to draw good quality conclusions from poorly collected evidence.

The evaluation process that we envisioned consists of two major stages. We first establish benchmarks for comparing effectiveness of multiple tools and work modes by conducting a series of end-to-end evaluations in a controlled environment. These evaluations must involve real analysts and realistic tasks and data

sources in order to produce reliable outcomes, i.e., under which conditions we can expect a particular level of analytic performance. Once the benchmarks are in place, we can attempt a hands-off automated and predictive evaluation (i.e., how is a tool "doing"). In order to accomplish this, we need to isolate intermediate *performance indicators*: automatically measurable variables of which values can be aligned with specific end outcomes. Such indicators may include: the number of data items shared, the number of data items retained per unit of time, the number of messages exchanged between analysts, the time spent searching vs. reviewing, etc.

The workshop reported here constituted only a "dry-run" of the first step in the above 2-stage process and its main purpose was to test the mechanics of the first phase evaluation before a longer-term study is attempted. The key objective was to see if the existing evaluation design, as well as the instruments and metrics are in fact adequate for measuring the effect of collaboration on the quality of analysis, and if not, what other or additional instruments and experiments may be needed. As it turns out, the workshop raised more questions than it answered; but it also exposed that the current evaluation design and metrics may not be sufficient.

The Analytical Tasks. Realistic analytical tasks were prepared with assistance from the sponsoring agencies. The tasks were selected and formulated to allow analysts to complete a report within the time limit set by the workshop organizers (2.5 hours per topic including time for report editing). The topics were selected to concern recent events of potential general interest, but with which the analysts were not likely to be very familiar. This last provision was included to minimize analysts' reliance on prior knowledge and thus to place more stress upon the system. We therefore selected 7 topics for evaluation that did not assume specialized knowledge on the part of the analysts but nonetheless displayed sufficient structural complexity (also reflected in the richness of the data available) to require both an analytic strategy and discipline in order to write meaningful reports. Here are titles of the selected topics:

(0) L-3 Communications Holdings, Inc
(1) Effect of Focused Vibrations on The Human Brain
(2) Tainted Chinese Food
(3) Risk of Cancer from Teflon-coated Products
(4) Artificial Reefs
(5) Honeybee Disappearance
(6) Chinese/Hong Kong IP Counterfeiting Operations

Each task was described using a brief narrative; below is an example task formulation:

Risk of cancer from Teflon-coated products

Please gather evidence and report on whether or not the use of Teflon-coated products (i.e., pans, pots, cookware) causes cancer in humans. List any reactions Teflon may cause when introduced by any means into the human body.

Describe the current state of research and evaluation on Teflon. List the organizations, government or otherwise, that are responsible for Teflon product safety, and evaluate the degree of bias in studies on this product. Add to your report any other relevant information.

Workshop Schedule. Table 1 shows the schematic training, work, and feedback schedule for the exercise. The first day was devoted to training and warm-up tasks to assess analysts' proficiency level with the system. The last day consisted of debriefs and focus groups with the participants.

Table 1. Schematic Workshop Schedule

	Day 1	**Days 2, 3 and 4**			**Day 5**
9-10 AM	Orientation	Analytic Topic 1	Analytic Topic 3	Analytic Topic 5	Final Debrief
10-11 AM	Training				
11-12 AM		Peer Evaluations & Questionnaires	Peer Evaluations & Questionnaires	Peer Evaluations & Questionnaires	
12-1 PM	Lunch break				
1-3 PM	Warm-up Task	Analytic Topic 2	Analytic Topic 4	Analytic Topic 6	
3-4 PM	Peer Evaluations & Questionnaires	Peer Evaluations & Questionnaires	Peer Evaluations & Questionnaires	Peer Evaluations & Questionnaires	
4-5 PM	Group Discussion	Group Discussion	Group Discussion	Group Discussion	

Work Modes. During the workshop analysts worked in groups of different sizes. Each group operated under one of the following conditions:

1. *Individual Work, No collaboration* (INC). Analysts were expected to prepare individual, independent best reports. No contact between analysts was allowed. This was the baseline against which other groups would be measured. At least 2 analysts were in this group.
2. *Individual Work, Tacit Collaboration* (ITC). Analysts were expected to prepare individual best reports, just like in INC; however, they were allowed to tacitly collaborate, with the system facilitating information sharing on as-needed basis. This work arrangement required no top-down coordination or management; analysts were free to share information and to communicate but each was expected to pursue an independent work strategy, and produce individual (though not necessarily independent) reports.
3. *Joint Work, Tacit and Open Collaboration* (JTC). Analysts were expected to collaborate by any means available. They were also required to prepare a single combined report, thus an up front division of work was normally assumed, with a leader/assembler elected at the outset (usually a senior group member). This work arrangement required time and asset management and coordination in order to create a successful product.

Table 2. The work modes rotation schedule

	JTC	**ITC**	**INC**
Task 1	ABCD	EF	GH
Task 2	EF	GHAB	CD
Task 3	GHAB	CD	EF
Task 4	CD	EFGH	AB
Task 5	EFGH	AB	CD
Task 6	GH	ABCD	EF
#reports	1	2 or 4	2

Group Size. We attempted to test effectiveness of collaboration in different size groups. Given the limited scale of our experiment we compared two group sizes across different collaborative settings: "small" groups of size 2 and "large" groups of size 4.[6] With 8 analysts (A–H) and 6 tasks (1–6), a rotating assignment schedule was established, as shown in Table 2. In this arrangement, most groups rotate through both collaboration modes (JTC, ITC); however, due to time limitations it wasn't possible to do for every group.

Training. We have developed training materials to introduce analysts to COLLANE. The training session was performed at the beginning of the workshop and included a tutorial, a hands-on tryout, and a warm-up task. The warm-up task was similar to the evaluation tasks but less complex. Its purpose was to allow the analysts to gain a degree of confidence in using the system, explain some "obvious" misunderstandings, and also to minimize the impact of unfamiliar technology on the first evaluation task. At the end of the warm-up task the analysts used the evaluation instruments, again to familiarize themselves with these aspects of the exercise. The entire training session took approximately 6 hours.

3.2 Preliminary Evaluation Results

COLLANE end-to-end performance was measured using the following metrics: (a) scoring of the final analytical reports for coverage, significance, organization, and other quality factors, and (b) questionnaires related to user satisfaction with the system and an assessment of their own performance using the technology, as well as other subjective factors such as workload perception. The report scoring was performed using the peer evaluation method (*cross-evaluation*) developed by the Albany team during the AQUAINT Project: in this approach all analysts

[6] Some aspects of collaborative group size are discussed in, among others, ([2]; [10])

act as a panel of (independent) judges producing multiple scores for each report, including their own. The questionnaires were developed to capture subjective assessment of particular aspects of the process that were not easily captured in the system logs. One of the questionnaires used was NASA TLX, which measures *individual perception of effort* put into the task. In addition, focus group interviews were conducted that solicited free form comments about participants' experience with the system and the exercise as a whole.

A number of objective performance indicators were also computed from the system logs; these included: the number and type of questions asked, the quantity of text snippets retrieved and retained into the report, relevance and utility of the returned answer elements, time needed to assemble various portions of the report, among others. These performance indicators, i.e., their content, order, and structure, are currently being correlated with end-to-end evaluation metrics for each of the task/user/mode settings in the exercise in order to isolate the performance characteristics that lead to a successful outcome (e.g., high-quality report), as well as those that may signal troubles. Once such performance indicators are isolated, we believe that they could be used to automatically monitor system performance outside of the controlled evaluation environment. This work is still ongoing and will be reported in a future publication.

Peer Evaluations. The cross-evaluation forms (Table 3) were used to facilitate scoring of the analytical reports based on their content and organization. Each participating analyst became an independent judge on the panel that reviewed and scored all reports generated during the session just completed. Since each session was devoted to a single analytical topic, the panel was, in effect, ranking the reports produced under different working conditions. An additional advantage of this method was that the judges are also the participants, and their scoring tends to reflect the experimental conditions: task difficulty, data availability, and time limits. Moreover, scoring assigned to own report (or own group report) provides additional cues on relative importance of certain information units.

The responses collected from cross-evaluation are tabulated and averaged over all judges, and then displayed as a bar chart. Figure 7 shows the cross evaluation results from one of the sessions (Artificial Reefs topic). In the chart, each bar represents an average cross-evaluation score (one for each of the 6 categories) assigned to each of the report coming out of this session. The tacit collaboration group (ITC) delivered 4 reports (4 leftmost bars in each bundle); the open collaboration group (JTC) delivered a joint report (5th bar), and the two single analysts (INC) delivered their own reports (rightmost 2 bars). We note that for this topic, the tacit collaboration group outperforms other groups in nearly all categories except for "coverage" (cat. 2). As we explain further below, this is not really the case since the reports in ITC group are often complementary, and thus should be judged as a "portfolio" rather than singly.

These results can be further averaged over all topics; however, in order to obtain a meaningful statistic we need to control for topic difficulty, analyst

Table 3. Report Cross-Evaluation Form

Please evaluate each report using the following criteria Use 5 point scale (1=awful, 5=great) and justify
1. Includes crucially important information Score: ①②③④⑤ Justification:
2. Has sufficient coverage Score: ①②③④⑤ Justification:
3. Avoids the irrelevant materials Score: ①②③④⑤ Justification:
4. Avoids redundant information Score: ①②③④⑤ Justification:
5. Is well organized Score: ①②③④⑤ Justification:
6. Overall rating of this report Score: ①②③④⑤ Justification:

experience and skill, as well as for the judge bias. To do so would require a significantly larger data sample than the current experiment provided. For this reason the results reported here should only be treated as indicative of some possible trends.

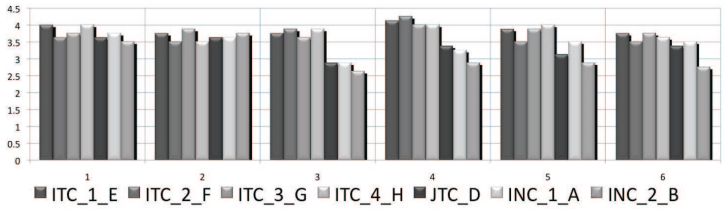

Fig. 7. Cross-evaluation scores for Artificial Reefs topic across all 6 criteria. The bars represent reports obtained in different work modes (ITC—tacit collaboration; JTC—open collaboration; INC—single mode).

Structured Questionnaires. Questionnaires were used to collect subjective opinions from the participants regarding their experience with COLLANE as well as with the various collaboration modes. We also sought analysts' opinions

about the evaluation process itself and whether they felt they had sufficient exposure to the new technology to make a judgment. All questionnaires were carefully designed to control for bias and to detect inconsistencies judgment (e.g., some questions were restated in a different form, etc.). Answers were recorded on a numerical scale allowing for computing scores that then could be averaged over all participants and all topics.

Post-session questionnaire

Post-session questionnaire consisted of 29 structured questions related to the task just completed by the analyst plus several free-form feedback questions. The questionnaire was administered immediately following each task. For collaborative tasks, each participant completed a separate questionnaire, thus multiple opinions were collected from collaborating teams. The questions sought analysts' assessment of the task itself (difficulty, appropriateness), specific features of the system (e.g., interface, speed, ease of use), and the work itself (collaboration, confidence, effort). All questions were structured so that they required answers on a 5-point Likert scale, which is normally presented as a set of "radio buttons" along with an intuitive scale, e.g., "not at all", "some", "a lot" (e.g., *How often did you use the visual interface?*) or "strongly disagree", "strongly agree" (e.g., *Did collaboration make analysis more efficient?*). The final free-form questions asked for comments on how to improve the existing system and how to make the work more efficient. The content of the questionnaire was adapted to each of the three work modes, i.e., questions concerning collaboration experience did not apply to analysts working singly. Below are a few sample questions from the beginning of post-session questionnaire:

1. How did this scenario compare to the tasks you perform at work?
(1: much less difficult 3: same 5: much more difficult)
2. How difficult was it to formulate questions for this task?
(1: much less difficult 3: same 5: much more difficult)
3. How confident were you about preparing a report for this task using COLLANE?
(1: not at all confident 3: confident 5: very confident)
4. How often did COLLANE respond to your questions with useful information?
(1: never 3: frequently 5: always)
5. How often did you find Rapid Results helpful?
(1: never 3: frequently 5: always)

Exit questionnaire

An exit questionnaire consisting of 32 questions was administered at the end of the workshop after all work sessions were completed. This questionnaire reprised some of the questions from the post-session questionnaires, now in a more general form and included additional questions about overall assessment of the system, the tasks, and the collaborative arrangements. We used the same general format of structured questions with responses collected on a 5-point Likert scale. Below are a few sample questions:

1. The COLLANE system allowed me to easily change my line of questioning.
(0: strongly disagree 5: strongly agree)
2. It was difficult to get the COLLANE system to do what I wanted it do?
(0: strongly disagree 5: strongly agree)

3. I easily understood the relationship between the question that I asked and the answer that the COLLANE system provided.
(0: strongly disagree 5: strongly agree)
4. The COLLANE system seriously slows down my process of finding information.
(0: strongly disagree 5: strongly agree)
5. The COLLANE system helps me find important information.
(0: strongly disagree 5: strongly agree)
6. The COLLANE system helped me think of new ways to search for information.
(0: strongly disagree 5: strongly agree)

We tallied the responses from post session and post workshop questionnaires in Figures 8 and 9, respectively. The charts in Figures 8 show that the analysts have generally found the COLLANE system satisfactory and the exercise realistic with task difficulty at the level typical for their professional experience. The scores within the 2.5 and 3.5 range represent the middle point where the analysts' expectations are being met, e.g., the task difficulty is compatible with their experience, the system response is in line with the technology they use at work, etc. We note that analysts were quite positive about their reports and expressed a fair amount of confidence in their results. We also note that current COLLANE information sharing and collaboration support capabilities are acceptable but clearly need more work to be truly satisfying. This suggests that COLLANE development is on the right track even though much work remains to be done.

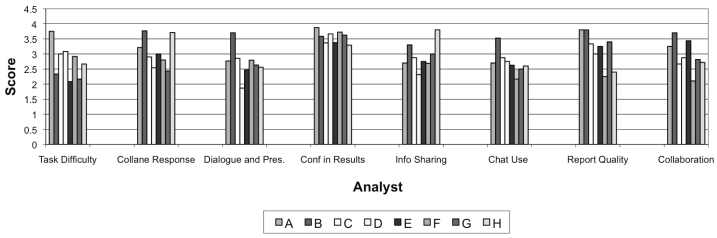

Fig. 8. Score averages from post-session questionnaires grouped into categories on 1-5 scale

Figure 9 shows average scores collected from the final questionnaire administered on the last day after all working sessions were completed. Again, we grouped the responses into several more intuitive categories. The assessment is very encouraging—analysts found COLLANE a promising technology while it is still a very preliminary prototype.

NASA TLX Questionnaire. NASA TLX instrument was used to assess subjective perception of workload during the task. Originally designed to test stress level for NASA astronauts, it has been adapted to analytic tasks with help of researchers from the National Institute of Standards and Technology. The revised questionnaire includes 6 categories of questions to rate analysts' experience with the task. Each question required a response on a 7-point scale,

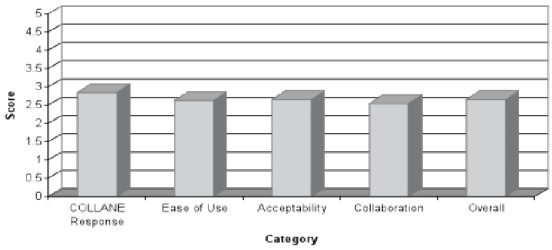

Fig. 9. Average scores from final questionnaire grouped into 5 categories

with 1 meaning 'little' and 7 standing for 'much'. The 6 categories were defined as follows:

1. *Mental demand*: to what degree does the task affect a user's attention, brain and focus
2. *Physical demand*: to what degree does the task affect a user's health, makes a user tired, etc.
3. *Temporal demand*: to what degree does the task take time that a user can't afford
4. *Performance*: to what degree is the task heavy or light in terms of workload
5. *Frustration*: to what degree does the task make a user unhappy or frustrated
6. *Effort*: how much effort did the user spend on the task

For all of the TLX categories a higher number assigned by the analyst corresponds to subjective perception of a higher cognitive workload. Figure 10 shows the averages from the TLX-1 questionnaires obtained from all analysts across all tasks. It is interesting to note how modes of collaboration make different demands upon the analyst: tacit collaboration requires more mental effort and time pressure, while open collaboration adds primarily physical effort. None of these differences are statistically significant given the small data sample.

System Logging. During the workshop, all analysts' activities were automatically logged into several data streams. These included all analytic actions performed on COLLANE user interface (questions asked by the user, responses and questions from the system, all browsing activities on the panels, access to source documents, etc.) as well as all copy and paste events from COLLANE answer space to the analyst report (assembled in a separate text document). In addition, for the collaborating analysts, their exchanges over the chat interface were recorded, including messages sent as well as the data items exchanged. All of these data streams were time stamped allowing for easy alignment and verification of each event.

The following is a partial list of key logged events. We should note that these capture "basic" analytic events that can be combined together to obtain more meaningful "analytic episodes" (e.g., exploring, drilling down, verifying,

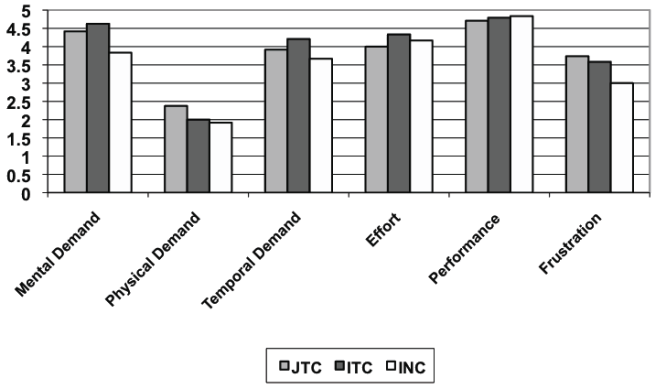

Fig. 10. TLX average scores for all sessions. The lower score is better except for the Performance category. The scale is 1 through 7.

etc.) which may lead to derivative utility-based metrics, e.g., time needed to assemble most of the information eventually included in the report, etc.

Examples of tracked user or system events:
Opening and closing an answer folder
Changing the relevancy of text passage
Displaying text through visual panel
Selecting an attribute to display on the visual panel
All dialogue between the users and the system
Bringing up a full document source
Text copied, and where it was copied from
Passages display and view
Browsing of the visual panel

One of the effect we were interested to note from the system logs was the degree and effectiveness of information sharing among the collaborating analysts, particularly in the tacit groups where the bulk of information sharing was automatically directed by the system. We wanted to see if apparently relevant but non-redundant information is correctly forwarded from one analyst to another, if it is being noticed by the recipient, and above all if it is being utilized in any way. We took copy events (into report draft) as evidence that a piece of information is being used; we also counted events where apparently viewed information (as evidenced by a passage display event) is ignored (i.e., not copied) or worse, it is labeled as non-relevant (e.g., by icon color change event). Figure 11 shows the effect of information sharing based on report usage across all three work settings. We note that the tacit collaboration group manages jointly to cover all key source citations.

We have also collected other quantitative information from the system logs. Some more interesting of these are reported below. For example, the graph

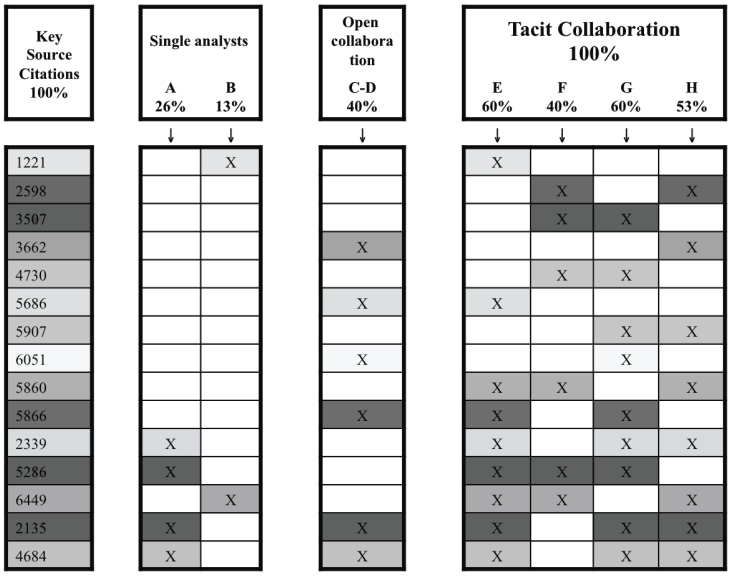

Fig. 11. Non-unique source citations among the analysts in Artificial Reefs task

in Figure 12 shows that on average, the analysts working in a collaborative setting required *less time to complete their tasks* than the analysts working alone. Specifically, analysts in the ITC mode with tacit information sharing completed their tasks faster than when working alone.

We need to note that while the open collaboration teams (JTC) completed their work faster than other analysts, the comparison is not straightforward. On the one hand, the analysts in open collaboration divide the task among themselves, which means that each analyst has a smaller problem to work with than the analysts in other groups; on the other hand, the JTC team needs to combine their partial reports, which requires extra time for assembly. It may be worth noting that time reduction also varies by the size of the JTC group: for topics 1, 3, and 5 (Vibrations, Teflon, and Honeybees), the JTC group had 4 analysts, thus the time reduction is more pronounced when compared to other work methods; for the other topics the JTC group had only 2 analysts and, as expected, the effect is lesser. In order to estimate the true time load in open collaboration, one would need to sum the times spent by each analyst. This would make the cost of open collaboration (measured per time unit) significantly higher than the cost of tacit collaboration. [7] Figure 13 shows that analysts in tacit collaboration (ITC) ask the system more questions, on average, than analysts working alone (INC). This is possibly a result of ITC analysts being

[7] There may, however, be an additional cost in tacit collaboration incurred because the output has not been integrated into one report.

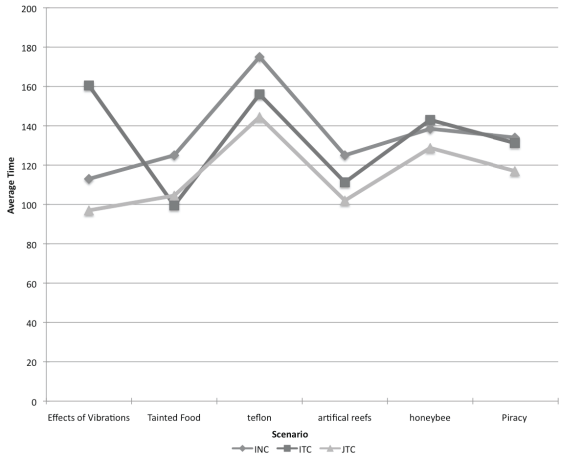

Fig. 12. Average time spent by individual analysts per topic for different modes of work (single, tacit, and open collaboration)

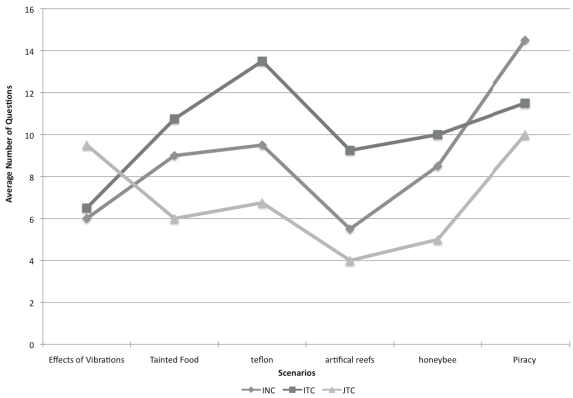

Fig. 13. Average number of questions asked by individual analysts working in different modes across all test topics

exposed to more evidence through tacit information sharing, and thus following more leads and researching their topics more thoroughly—a clearly desirable effect. We also note that tacitly collaborating analysts, while asking more questions, spent less time on their tasks than when working alone (cf. Fig. 12 vs. Fig. 13). This seems to indicate that analysts in tacit collaboration work faster and do more than in other work modes—another highly desirable effect. As before, we can't directly compare analysts' performance in open collaboration (JTC); while they asked fewer questions, this is most likely an effect of task

subdivision, which does not occur in either ITC or INC. Nonetheless, it may indicate that JTC is a less productive form of collaboration than ITC.[8]

The primary source of citations in the report was the passage text encapsulated into the frames in the visual display (approx. 64%), with additional 11% coming from rapid results and the answer folders. In other words, 75% of cited material was selected from the passages directly offered by the system as relevant. The remaining 25% came from other parts of the retrieved documents, i.e., other passages than those explicitly displayed on the interface. This attests to the high-degree of precision of the COLLANE question answering component (Figure 14).

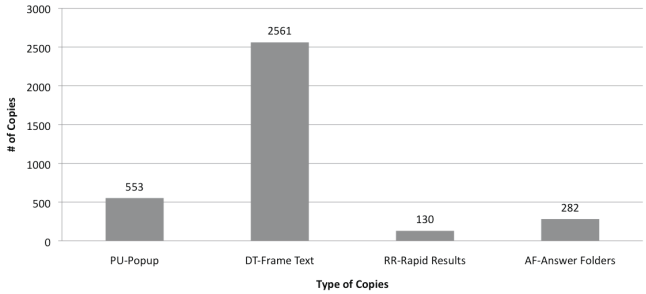

Fig. 14. Sources of citations in the reports

Finally, in Figure 15 we show selected performance statistics by the level of analytic experience. These results are averages over all sessions and work methods, and thus must be viewed only as illustrative (as discussed before, averaging scores in collaborative teams is not appropriate). Nonetheless, we note that while the most experienced analysts were in fact most effective and efficient in asking the right questions (their rate of productive questions[9] is very high at 71%), this effectiveness does not seem to translate into the higher report quality, as evidenced by the average cross-evaluation scores assigned to final reports (Scores columns in the chart).[10]

3.3 Summary of Evaluation Results

The preliminary results from this study suggest that COLLANE-supported tacit collaboration has the potential to produce significantly higher quality intelligence

[8] There may be an inclination to take less responsibility, or take a more passive role, when the task has been divided up, or when there is a designated task head who will lead the integration of the report.

[9] Productive questions (Good Questions in the chart) are defined as these that return at least one passage (citation) that is saved into the report.

[10] This may well be just a side effect of the current experiment, related to the selection of test topics. Further analysis, over a larger test sample is needed.

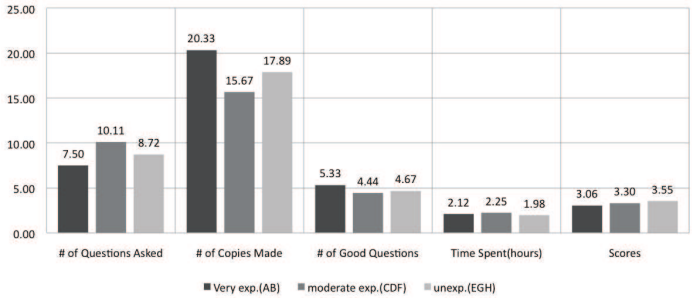

Fig. 15. Analytic effectiveness and efficiency by experience

than would be possible by analysts working alone or in open collaboration groups. The results also show that to properly measure the impact of collaboration will require new evaluation techniques that go beyond the current metrics. This early prototype of COLLANE was well received by the analysts; nonetheless, more advanced research is required to exploit the full potential of collaborative technology on the analytic process. The workshop provided findings at three levels, as follows:

1. COLLANE Information Sharing and Tacit Collaboration are beneficial
   - ***Information sharing in COLLANE improves the analytic process.*** Analysts working in collaborative teams (open and tacit) are exposed to more topic-relevant information than analysts working alone. Therefore, we may infer that the collaborating analysts are using more evidence to arrive at more informed conclusions (example: Fig. 11).
   - ***The quality of the reported intelligence improves.*** Qualitative assessment of the reports prepared by the collaborating teams suggests that better and more conclusions are drawn by collaborating analysts than by analysts working alone (ex. Fig. 7).
   - ***Current benefits of collaboration via COLLANE are only preliminary.*** In order to obtain full benefits of tacit collaboration, COLLANE information sharing must be advanced to knowledge sharing capabilities.
2. Effects of Collaboration on the Analytic Process are positive
   - ***Collaboration benefits the analytic process*** but its effects vary between different forms. We found evidence that tacit collaboration is useful, but there was insufficient information to understand the effects of open collaboration.
   - ***Tacit collaboration introduces a game-like competitive element.*** Analysts working in tacit collaboration mode tend to pursue alternative or complementary interpretations of evidence. Unlike in the open collaboration, they are not being driven into consensus.

   - ***Tacit collaboration improves analytic productivity*** by inducing analysts to do more useful work per time unit. This is likely caused by increased information sharing and an element of competition. Tacit collaboration allows the analysts to benefit from the experience of others on their team without necessarily slowing them.
   - ***Tacit collaboration helps to expose alternative interpretations*** of the available evidence; the outcome is a portfolio of reports that facilitates the survival of sound alternatives.
3. New Evaluation Methodology is required to measure the full effects of collaboration
   - ***Current evaluation methodology is inadequate*** because it is geared towards a single output of the analytic process (report quality) and does not support the evaluation of multiple outcomes of collaboration.
   - ***Evaluation of report quality is insufficient*** because there is not enough time to do it in the current design and also there is no independent assessment of the output by an external judge panel. The latter is particularly important for evaluating the output of tacit collaboration.
   - ***New metrics are required*** to assess the relative value of a multifaceted portfolio of reports covering a complex intelligence problem.
   - ***Revised evaluation design is required*** in order to control for confounding factors such as level of experience, subject matter expertise, and analytic skills of the participants. In addition, we need ways of measuring effects of differently skilled participants on team performance.

3.4 Challenges

Not surprisingly, we found a number of challenges that would need to be addressed in future evaluations. Some of these challenges were already signaled in the preceding sections. Here we summarize them briefly:

- *Information sharing vs. knowledge sharing.* While information sharing currently facilitated by COLLANE is clearly beneficial, further improvements are expected by supporting knowledge sharing, i.e., the source information along with questions, annotations, and exploratory metadata left by the analysts.
- *Collaboration vs. competition.* Tacit collaboration seems an effective way of improving analytic effectiveness, but it also involves an element of competition, which needs to be taken into account. It requires further study to determine if this finding holds across further testing, and to determine if there are possible negative effects as well.
- *Collection vs. judgment.* Increased information sharing helps analysts to collect more supporting evidence but we need to provide tools to convert more information into better judgments.
- *"Correctness" of conclusion vs. soundness of argument.* We are interested in supporting sound arguments based on the available evidence, not simply the "correct" conclusion, which may be consistent with others' viewpoints.

- *Process quality vs. results quality.* We are interested in measuring both the process quality and the results quality, as well as in optimizing the connection between the two.
- *View rotations.* The current process does not support analysts' viewing each other's progress (e.g., evolving hypotheses); however such capability may be highly beneficial. Analysts expressed interest in being able to peek over "each other shoulders".

We have also identified challenges related to the mechanics of preparing an evaluation exercise. These include issues such as data collection, problem preparation, recruiting analysts, etc. Here are some of the key issues:

- *Data preparation*: The web mining method which depends on harvesting thousands of documents from the web through a series of rapid searches may not be adequate for creating data collections that can support analysis of highly complex topics; there is simply no guarantee that all relevant (and related but not relevant) aspects will be included. Potential remedies include human-in-the-loop interactive mining, mining more data, or using the open sources (e.g., internet). This last option has been frequently invoked by the participants; however, it raises a number of issues including stability of the experiment.
- *Time and scope*: The compressed nature of the experiment does not align well with typical analytical experience where analysts work on multiple topics but spend considerably more "clock" time on each. Analysts suggested that they should have an entire day to research a topic.
- *Access to the Internet*: Access to the open internet was not provided so as to maintain a controlled experimental environment. Nonetheless, analysts felt this limited their options too much, especially when dealing with unfamiliar and complex topics.
- *Access to specific data resources*: Some of the information needs raised by the tasks could be most naturally satisfied by searching specific data repositories (e.g., Government regulations on artificial reefs). Analysts thought that doing open-ended search for information that is readily available elsewhere made some aspects of the exercise unrealistic. We need to find a better way to balance the experimental needs with the realism of the evaluation exercise.
- *Access to other tools*: In addition to the above, access to other analytic tools was occasionally called for. Specifically, tools for organizing collected information by event date or release date was requested as an essential management tool.

4 Conclusions

In this chapter we described an advanced analytic system COLLANE and a process of conducting a task oriented evaluation with real users. COLLANE, which

is in an early prototype stage, has been designed specifically to facilitate and support tacit and open collaboration among a group of analysts working on complex information problems. The preliminary evaluation described in this paper has led to a number of (indicative) observations, all of which need to be confirmed through further research:

1. Tacit collaboration is an effective means of improving analytic processes;
2. Tacit collaboration leads to better quality results;
3. Tacit collaboration does not drive analysts to a consensus; instead it exposes alternative approaches to complex problems;
4. Collaborative analysis is more efficient but also more demanding than working alone;
5. New metrics are required to adequately measure the full benefits of collaboration.
6. Information sharing may need to extend towards exchange of partially structured knowledge to further enhance the power of tacit collaboration.

Acknowledgements

We would like to thank Dr. Emile Morse for her assistance with the workshop planning and selection of evaluation instruments. LCDR Joseph Henriquez was instrumental in assembling the team of analysts who participated in the experiment and assisted in selection of the analytical tasks. This report is based on work supported by the IARPA CASE Program under the contract to SUNY Albany.

References

1. Allen, J. and Core, M.: Draft of DAMSL: Dialog Act Markup in Several Layers `http://www.cs.rochester.edu/research/cisd/resources/damsl/` (1997)
2. Avouris, Nikolas; Meletis Margaritis, Vassilis Komis: The effect of group size in synchronous collaborative problem solving. In *Proceedings AACE Conf. EDMEDIA* (2004)
3. Dillenbourg, Pierre and David Traum Sharing solutions: persistence and grounding in multi-modal collaborative problem solving. In *Journal of the Learning Sciences.* (2005)
4. Hardy, H., et al.: The AMITIES System: Data Driven Strategies for an Automated Dialogue In *Speech Communication* 48, Elsevier. 354-373. (2006)
5. Hardy, H., V. Kanchakouskaya, and T. Strzalkowski: Automatic Event Classification Using Surface Text Features. In *Proceedings of the AAAI Workshop on Event Extraction and Synthesis* (2006)
6. Kelly, Diane, Nina Wacholder, Robert Rittman, Ying Sun, Paul Kantor, Sharon Small, and Tomek Strzalkowski: Using Interview Data to Identify Evaluation Criteria for Interactive, Analytical Question-Answering Systems. In *Journal of the American Society for Information Science and Technology* (JASIST) (2007)
7. Kirchhoff, K.o and M. Ostendorf: Directions for multi-party human-computer interaction research. In *HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing* (2003)

8. Lemon, O., P. Parikh, and S. Peters: Probabilistic Dialogue Modeling. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, Philadelphia. (2002)
9. Morse, E.: An Investigation of Evaluation Metrics for Analytic Question Answering. In *Proceedings of AQUAINT Phase II*,6-month PI Meeting, Tampa. (2004)
10. Ryall, Kathy; Clifton Forlines, Chia Shen, Meredith Ringel Morris: Exploring the Effects of Group Size and Table Size on Interactions with Tabletop Shared-Display Groupware. In *CSCW '04*, Chicago, Illinois, USA. (2004)
11. Small, Sharon; Tomek Strzalkowski, Hilda Hardy, Nick Webb, and Boris Yamrom: HITIQA: High-Quality Intelligence through Interactive Question Answering. in *Journal of Natural Language Engineering, Cambridge.* (to appear 2008)
12. Small, Sharon: An Effective Implementation of Analytical Question Answering. *Doctoral Dissertation*, Computer Science, SUNY Albany. (2007)
13. Strzalkowski, T. et al.: Collaborative Analytical Workshop with COLLANE, Preliminary Report. *Submitted to IARPA.* (2007)
14. Strzalkowski, T. and S. Harabagiu: Advances in Open-Domain Question Answering. *Springer.* (2006)
15. Strzalkowski, T., S. Small, H. Hardy, B. Yamrom, T. Liu, P. Kantor, K.B. Ng, N. Wacholder: HITIQA: A Question Answering Analytical Tool. In *Intenational Conference On Intelligence Analysis.* (2005)
16. Strzalkowski, T., S. Small, S. Taylor, B.A. Lipetz, H. Hardy, N. Webb: Analytical Workshop with HITIQA. *A preliminary Report to IARPA.* (2006)
17. Strzalkowski, Tomek, Sharon Small, Hilda Hardy, Ting Liu, Sean Ryan, Nobuyuki Shimizu and Min Wu: Question Answering as Dialogue with Data. In *Advances in Open-Domain Question Answering,* pp. 149–188 Springer (2006)
18. Traum, D R., CF. Andersen, W Chong, et al.: Representations of Dialogue State for Domain and Task Independent Meta-Dialogue. *Electron. Trans. Artif. Intell.* 3(D): 125-152. (1999)
19. Traum, D R., J. Rickel: Embodied agents for multi-party dialogue in immersive virtual worlds. *AAMAS,* 766-773. (2002)
20. Wacholder, N, P. Kantor, S. Small, T. Strzalkowski, D. Kelly, R. Rittman, S. Ryan, R. Salkin: Evaluation of the HITIQA Analysts' Workshops. *Final Report.* (2003)
21. Wacholder, Nina, Diane Kelly, Paul Kantor, Robert Rittman, Ying Sun, Bing Bai, Sharon Small, Boris Yamrom, and Tomek Strzalkowski: A Model for Quantitative Evaluation of an End-to-end Question Answering System. In *Journal of the American Society for Information Science and Technology* (JASIST). (2006)
22. Webb, N., M. Hepple and Y. Wilks: Dialogue Act Classification based on Intra-Utterance Features. In *Proceedings of AAAI workshop on spoken language understanding.* (2005)
23. Webb, Nick; Hilda Hardy, Cristian Ursu, Min Wu, Yorick Wilks, and Tomek Strzalkowski: Data-Driven Language Understanding for Spoken Language Dialogue. In *Proceedings of the AAAI Workshop on Spoken Language Understanding.* Pittsburgh, PA. (2005)