

Question Generation and Answering Systems R&D for Technology-Enabled Learning Systems Research Roadmap

This is one in a series of component roadmaps developed by the Learning Federation (www.thelearningfederation.org). Five roadmaps have been developed in all, focusing on the following topics:

- Instructional Design: Using Simulations and Games in Learning
- Question Generation and Answering Systems
- User Modeling and Assessment
- Building Simulations and Virtual Environments
- Integration Tools for Building and Maintaining Advanced Learning Systems

The Learning Federation roadmaps are designed to provide a vision of where we can go with adequate investment in learning science and technology R&D and a detailed research plan that can achieve that vision. Our goal is to catalyze a partnership joining companies, universities, government agencies and private foundations to execute the research plan and make possible radically improved approaches to teaching and learning enabled by information technology.

The roadmaps have been developed through an iterative process that includes: literature reviews; interviews with researchers and practitioners; and workshops that convene experts from educational institutions, government, diverse industry representatives, and software publishers. Draft documents have been circulated widely in the research and development communities and revised in response to the valuable feedback so generously provided. We will be convening national leaders to determine the appropriate scale and management of the overall research and development effort.

Table of Contents

1	OVERVIEW AND PURPOSE	3
2	PROCESS USED TO DEVELOP THE ROADMAP	3
3	INTRODUCTION	4
4	ISSUES AND TECHNICAL CONSIDERATIONS	6
4.1	KNOWLEDGE NEEDED TO BUILD HARDWARE AND FUNCTIONING SOFTWARE TOOLS	6
4.2	CLASSIFICATION OF LEARNING CONTENT AND LEARNING OBJECTIVES	7
4.3	CLASSIFICATION OF QUESTIONS	12
4.4	MEASURING PROGRESS	14
4.5	MAJOR DESIGN OBJECTIVES OF TOOLS	17
5	RESEARCH ROADMAP	18
5.1	LEARNING ENVIRONMENTS THAT STIMULATE LEARNER QUESTIONS, AND QUESTIONS THAT STIMULATE LEARNING	18
5.2	INTERFACE FOR LEARNERS TO ASK QUESTIONS.....	24
5.3	INTERFACES FOR ASKING LEARNERS QUESTIONS AND ANSWERING LEARNER QUESTIONS	30
	MILESTONES	32
5.4	COMPREHENDING AND ANSWERING LEARNER QUESTIONS	33
5.5	MECHANISMS FOR INTERPRETING LEARNER ANSWERS.....	36
5.6	COLLABORATIVE MIXED-INITIATIVE DIALOG	40
5.7	USER MODELING.....	45
	MILESTONES	47
5.8	INCORPORATING LEARNING PEDAGOGY	48
5.9	LINKING PEOPLE INTO THE QG&A SYSTEMS.....	50
6	LAUNCHING AND STIMULATING THE QG&A RESEARCH AGENDA	50
7	REFERENCES	51

1 Overview and Purpose

This document is a research roadmap for developing technologies that can facilitate question asking and answering in learning environments. Learning often improves when students are stimulated to ask questions and when there are facilities for receiving relevant, correct, and informative answers. However, there is a need to identify (a) the conditions in which question-based learning is effective, (b) the landscape of questions that learners ask and should ask, (c) the alternative facilities for having good answers delivered or collaboratively constructed, (d) the technologies that must be developed to support question-based learning, and (e) what a “good answer” is for whom and under what circumstances. A broad range of learning environments must be considered: classrooms, laboratories, electronic textbooks, one-on-one tutoring, collaborative work, helpdesks, and job aids.

New technologies can make learning more productive, compelling, personal and accessible. The big question is how to structure the field of learning and education from the standpoint of question generation and answering. How can we move to the next level? A plan needs to be developed to integrate technological advances in learning. This roadmap provides such a plan and suggests a research agenda for the next 3, 5, and 10 years.

The focus will be on education and training at the level of postsecondary education. This includes courses at 2-year and 4-year universities and colleges, as well as lifelong learning experiences in business, industry, and the government. There will be an emphasis on science, mathematics, engineering, and technology.

2 Process Used to Develop the Roadmap

The roadmap was developed in a collaborative effort that brought research and development leaders together across disciplines, from both the public and private sectors, to identify promising research that integrates question generation and answering capabilities into learning systems. The roadmap was developed as follows:

- A review of recent research on questioning and answer systems from multiple perspectives (psychology, cognitive science, computational linguistics, information systems design) was conducted by the Federation of American Scientists;
- Interviews were conducted with relevant researchers, program managers, and industry representatives;
- Research priorities were developed from the collective expertise of representatives from companies, universities, government research facilities, and others with unique expertise.

- A workshop of acknowledged experts, researchers, developers, and implementers of question generation and answering systems to evaluate, discuss and contribute to the final report.
- Through an iterative process research tasks were identified and 3, 5, and 10 year milestones were developed;
- Discussions from the workshop were synthesized into the roadmap report which was disseminated to workshop participants and other experts. Comments on the report were incorporated into this final report. The list of workshop participants and report reviewers is provided in Appendix A.

3 Introduction

Advanced learning systems become more sophisticated when they stimulate learner questions and answer the questions that learners ask. Throughout history, great teachers have sought to create learning environments that encourage questions and provide thoughtful, helpful responses. Those who have allegedly used Socratic teaching styles, for example, have attempted to guide their students through a process of interrogation in order to help them identify the limits of their own knowledge, to discover misconceptions, and to achieve genuine self-knowledge. Researchers in cognitive science and education have advocated and have shown learning benefits for environments that encourage students to generate questions (Beck, McKeown, Hamilton, & Kucan, 1997; Dillon, 1988; Graesser, Langston, & Baggett, 1993; King, 1994; Miyake & Norman, 1979; Pressley & Forest-Pressley, 1985; Schank, 1999). Question generation is understood to play a central role in learning because it both reflects and promotes active learning and construction of knowledge (Bransford, Goldman, & Vye, 1991; Brown, 1988; Graesser & Wisher, 2002; Otero & Graesser, 2001; Papert, 1980; Scardamalia & Bereiter, 1985).

A key challenge to researchers and practitioners alike is to find ways to facilitate inquiry by taking advantage of the benefits offered by emerging technologies. That is, how can we use what we know about question generation and question answering to design learning environments that guide learners to actively construct knowledge? One obvious way to do this is to have a computer facility that is capable of answering student questions whenever they ask them. The system would formulate answers in a fashion that uses the specific pedagogical theory deemed most appropriate for the learner and subject. However, the computer facility would be used only if it delivers quick, correct, relevant, and informative answers. As a longer-term goal, one could imagine even more sophisticated facilities that diagnose student problems and provide help *before* the question is asked. That would require detailed learner profiling that not only keeps track of general capabilities and aptitudes of the learner, but also details about the history of learning episodes. But in addition to question answering facilities, facilities are needed for stimulating learner questions. It is well documented that student questions are rare in most learning environments (Dillon, 1988; Graesser & Person, 1994), so we need to identify learning situations that stimulate questions, such as challenges, contradictions,

and obstacles to important goals (Graesser & Olde, in press). We know that learning improves when learners are taught how to ask good questions, either through direct instructions on question asking (King, 1989, 1994; Rosenshine, Meister, & Chapman, 1996) or by a person or computer that models good question asking skills (Craig, Driscoll, & Gholson, 2002; Palincsar & Brown, 1984). The research agenda includes methods for increasing the frequency and quality of questions, as well as methods for delivering answers to learner questions.

The development of effective question generation and answering (QG&A) facilities requires a detailed understanding of a complex multifaceted system. The system has many components. The components include, but are not limited to:

- the cultures of the society of learners
- the teachers' pedagogical theories
- the goals of the learner
- concrete learning scenarios
- the potentials and limitations of available technologies
- the curricula mandated by schools systems and government agencies

The research agenda requires a coordinated multidisciplinary effort from education, psychology, cognitive science, communication, human-computer interaction, software engineering and design, information science, computational linguistics, statistics, subject matter experts, and nonacademic sectors of society. The research priorities in this report were developed from the collective expertise of representatives from universities, companies, and government research centers. This final report will serve as a comprehensive and strategic view of the field which researchers, industry, and funding agencies can use as a guide for developing and applying QG&A systems in learning environments.

The process of implementing effective QG&A facilities into technologically-based learning systems requires the integration of recent research and practice from multiple disciplines in both academia and industry. Unfortunately, rapid advances in these fields, as well as the traditional isolation of academic disciplines, make tracking and integrating related findings difficult. Relevant commercial products are often developed without knowledge of ongoing research, and university research groups frequently do not incorporate developments from industry. This disconnect hinders the rate and quality of innovation and the ability of education systems to maximize the benefits of computer-use for teaching and learning. This research roadmap fills a critical need to raise awareness of QG&A mechanisms in learning environments. Stakeholders need to have a coordinated understanding of the relevant research results, computational tools, on-going programs and projects across research disciplines, industry efforts, and government funding organizations. This roadmap will hopefully encourage dialog and partnerships to leverage gains from one field to other fields.

4 ISSUES AND TECHNICAL CONSIDERATIONS

This section identifies some issues and technical considerations that need to be addressed in planning the roadmap and measuring progress toward meeting our goals.

4.1 Knowledge needed to build hardware and functioning software tools

The research objective is to develop foundational knowledge that can be used to build functioning software tools that support QG&A facilities in learning environments. Some of this technical knowledge incorporates the latest advances in hardware, software engineering and computer science, whereas other knowledge builds on research and practice in education and cognitive science. QG&A facilities for education will not be effective if there is disconnect between technology and the learning process. We also need to understand how to integrate the two.

A full range of computational environments is anticipated when filling out the research roadmap. The learning environments will in principle accommodate desktop applications, the Internet, telecommunication, wireless communication, handheld devices, environmental sensors, and alternative input/output media. The design of the components will be constrained by what is known about question generation and answering in learning environments. Collins, Neville, and Bielaczyc (2000) identified many of the strengths and limitations of different types of media and multimedia in the design of learning environments that range from textbooks to synchronous communication on the web. Appropriate selection of media and multimedia depend on the subject matter, the learning goals, and the inherent capabilities of the various media/multimedia.

For each technology component (i.e., hardware, software, media), there needs to be a careful analysis of how it would be used by the learner. How long does it take for the learner to learn how to use it? Once learned, what are the task completion times and error rates when learners complete representative benchmark learning tasks? What is the attrition (dropout) rate for using a technology component after a person first masters it? The field of human-computer interaction has developed analytical schemes for systematically analyzing the use of specific technological artifacts. The GOMS model (Card, Moran, & Newell, 1983) simulates task completion times and error rates when specific technological artifacts are used by adults who have mastered them to the point of automaticity (overlearning). Carroll (2000; Carroll & Rosson, 1992) has developed a scenario-based approach to human-computer interaction which tracks the full life-cycle of the use of a technological artifact in an organization. The use of an artifact very much depends on the practices, social organization and culture of the users in addition to the purely cognitive constraints. One obvious conclusion from the 25-year history of human computer interaction is that the QG&A facilities developed in our research roadmap will require a careful cognitive and social task analysis from start to finish. Otherwise, the hardware components will collect dust, the software and media products will have no

contact with human minds, and those who manage large training and education budgets will lose interest.

The development of functioning software tools is expensive, so it is important to minimize costs. One way to accomplish this economic objective is to develop software tools that can be adapted for use by many different operating systems and repositories of learning content. For example, in 2000 the Department of Defense launched an ADL/SCORM initiative (www.adlnet.org), which stands for Advanced Distributed Learning/Shareable Content Object Reference Model. These standards are being advanced by the DoD, IEEE and IMS. The goal is to have course content be packaged and formatted so that it can be shared with different learning management systems. Smaller units of content (a paragraph, a graph, a simulation) can be reused by an unlimited number of lesson planners who develop courseware for specific purposes. Ideally, the unique features of a learner's profile will guide the learning management system to select and dynamically order the smaller units of content in a fashion that is uniquely tailored to the individual learner. This approach should radically decrease costs, but at the same time allow the development of complex learning modules. It becomes economically feasible to invest several million dollars on a learning project if the content (both static and dynamic) can be shared with thousands or millions of learners.

4.2 Classification of learning content and learning objectives

One truism in education circles is that the effectiveness of a learning environment depends on the learning objectives and the learning content. This claim is informative, both in science and in practice, only when there is a principled way of classifying learning objectives and learning content. As one might imagine, researchers in education and cognitive science have proposed a large number of classification schemes. One goal for the research roadmap is to judiciously select a landscape of learning objectives and content, with the explicit acknowledgement that this landscape will evolve over time.

One contrast that is frequently made is between shallow and deep knowledge (Bloom, 1956; Bransford et al., 1991; Brown & Campione, 1996; Chi, deLeeuw, Chiu, & LaVanher, 1994; Kieras & Bovair, 1984). There is no ironclad agreement on what characteristics precisely differentiate deep from shallow knowledge, but cognitive scientists have offered a number of likely candidates. According to Graesser, Otero, & Leon (2002), shallow knowledge consists of explicitly mentioned ideas in the learning material that refer to lists of concepts, a handful of simple facts or properties of each concept, simple definitions of key terms, and major steps in procedures. Deep knowledge consists of coherent explanations of the material that help the learner generate inferences, solve problems, make decisions, integrate ideas, synthesize new ideas, decompose ideas into subparts, design systems, forecast future occurrences in a system, and apply knowledge to practical situations. Deep knowledge is presumably needed to articulate and manipulate symbols, formal expressions, and quantities, although some individuals can successfully do this without deep mastery. In the case of procedural learning and problem solving, students can learn to manipulate and even generate formal

representations in the absence of a deep understanding of the formal structures or their underlying semantics. Because their knowledge is shallow, they will fail to execute these skills when appropriate outside the learning environment, will fail to transfer the knowledge to similar situations, and will not be able to interpret and use the outcome of the procedures. Most cognitive scientists agree that deep knowledge is essential for handling challenges and obstacles because there is a need to understand how mechanisms work and to generate and implement novel plans. Explanations are central to deep knowledge, whether the explanations consist of logical justifications, causal networks, or goal-plan-action hierarchies. It is well documented that the construction of coherent explanations is a robust predictor of an adult's ability to learn technical material from written texts (Chi et al., 1994; Cote, Goldman, & Saul, 1998; Webb, Troper, & Fall, 1995).

The cognitive representations of texts and pictures can be segregated into the levels of explicit information, mental models, and pragmatic interaction. The explicit information preserves the wording, syntax, and semantic content of the material that is directly presented. The mental model is the referential content of what explicit material is about. For everyday devices, for example, this would include: the components of the electronic or mechanical system, the spatial arrangement of components, the causal chain of events when the system successfully unfolds, the mechanisms that explain each causal step, the functions of the device and device components, and the plans of agents who manipulate the system for various purposes. The pragmatic communication level specifies the main messages that the author or graphics designer is trying to convey to the learner. More inferences are needed as one moves from the explicit information to the mental models to the pragmatic communication levels. Thus, a very different array of instructional events would be entailed.

It is possible to get quite specific in classifying the content of the material to be learned. One can classify knowledge into different branches of mathematics, sciences, technology, humanities, and arts according to a rich multilayered taxonomy. Researchers in artificial intelligence and cognitive science have spent three decades attempting to represent and classify knowledge according to more principled constraints (Lehmann, 1992; Lenat, 1995; Schank, 1999). As discussed later, the classification of knowledge has nontrivial consequences for the proposed roadmap because specific question categories are associated with particular types of knowledge representation (Graesser & Wisher, 2002). Some example categories of knowledge representation are enumerated in Table 1, but it is yet to be decided what classification scheme would suite the purposes of the GG&A roadmap.

Table 1: Categories of knowledge representation.

Agents and entities. Organized sets of people, organizations, countries, and entities.

Class inclusion. One concept is a subtype or subclass of another concept.

Spatial layout. Spatial relations among regions and entities in regions.

Compositional structures. Components have subparts and subcomponents.

Procedures & plans. A sequence of steps/actions in a procedure accomplishes a goal.

Causal chains & networks. An event is caused by a sequence of events and enabling states.

Others. Property descriptions, quantitative specifications, rules, mental states of agents.

Cognitive processes also vary in difficulty. Table 2 lists the major types of cognitive processes that are relevant to an analysis of questions (Standards of the Army Training Support Center, 2000; Bloom, 1956; Snow, 2002). According to Bloom's taxonomy of cognitive objectives, the cognitive processes with higher numbers are more difficult and require greater depth. Recognition and recall are the easiest, comprehension is intermediate, and classes 4-7 are the most difficult. Once again, this classification serves as a first-cut analysis and is not etched in stone. Whatever classification scheme is adopted, it is important for the designer of the learning environment to declare the learning objectives that guide the design of the educational software.

Table 2: Types of cognitive processes that are relevant to questions.

- (1) **Recognition.** The process of verbatim identification of specific content (e.g., terms, facts, rules, methods, principles, procedures, objects) that was explicitly mentioned in the learning material.
 - (2) **Recall.** The process of actively retrieving from memory and producing content that was explicitly mentioned in the learning material.
 - (3) **Comprehension.** Demonstrating understanding of the learning material at the mental model level by generating inferences, interpreting, paraphrasing, translating, explaining, or summarizing information.
 - (4) **Application.** The process of applying knowledge extracted from the learning material to a problem, situation, or case (fictitious or real-world) that was not explicitly expressed in the learning material.
 - (5) **Analysis.** The process of decomposing elements and linking relationships between elements.
 - (6) **Synthesis.** The processing assembling new patterns and structures, such as constructing a novel solution to a problem or composing a novel message to an audience.
 - (7) **Evaluation.** The process of judging the value or effectiveness of a process, procedure, or entity, according to some criteria and standards.
-

A somewhat different approach to analyzing learning content and objectives appeals to specific cases, problems, projects, or scenarios of use (Carroll, 2000). The term scenario is adopted here as a covering term for a cluster of similar constructs. A scenario is a more complex situated activity that takes a longer time to achieve, that recruits several cognitive representations and processes, that is anchored in a specific spatial and social setting, and that is guided by a number of goals that one or more agents want to achieve. Consider, for example, the following scenarios that would presumably have rather different learning environments.

(1) **An electronic textbook on general science.** An encyclopedia on general science is available on the web and desktop, with pictures and illustrations, for learners to peruse in hypertext/hypermedia. The learner can ask questions about concepts in the textbook and receive answers. This is available as supplementary material for any science course at the post-secondary level (2-year and 4-year colleges and universities) and lifelong science, math, engineering and technology education. It is also available in industrial settings for workforce development needs. The electronic textbook serves a broad set of needs, covers a broad set of topics, and covers both shallow knowledge and rudiments of deep knowledge. It does not excel in handling deeper cognitive processes (levels 4-7 of Table 2) and practical procedures that accomplish specific goals.

(2) **Newtonian physics problems with computer simulation and hypothetical reasoning.** Learners interact with an intelligent tutoring system on the web or desktop, where they learn about principles of Newtonian physics. They receive word problems such as: “A lightweight car and a massive truck have a head-on collision. On which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion? Explain why.” The learner holds a conversation with the computer tutor in natural language, as the tutor and learner collaboratively construct an answer to this deep-reasoning question (see Graesser, VanLehn, Rose, Jordan, & Harter, 2001 for descriptions of AutoTutor and Why/Atlas intelligent tutoring systems with natural language dialog). There is also a microworld with vehicles simulating the collision. Learners can manipulate a fixed set of parameters of the collision (e.g., the speed and mass of each vehicle, air resistance) and see what happens. Learners then describe what they see and explain why the collisions occur the way they do. The important categories of questions have stems such as *why*, *what if*, and *what if not*, because explanation and hypothetical reasoning are important objectives. Learners are free to ask questions in natural language and the computer supplies answers that are sensitive to the history of the dialog and the acts of manipulating parameters of the simulation. The answers are dynamically constructed by (a) interpreting the question (natural language understanding), (b) searching an electronic textbook for paragraphs with answers (information retrieval), (c) selecting information from fetched paragraphs for relevant information (information extraction) and (d) formulating responses (natural language generation).

(3) **Classroom analysis of medical problems.** The teacher and classroom of students are presented a case in which a patient has a hearing problem. The case is displayed on the web for everyone in the classroom to view. The students recommend methods of searching the web for information that might help solve the ear problem (e.g., *Google* searches and web sites on the topic of the ear). The teacher pursues some of the recommendations. The teacher periodically asks the students questions, which they simultaneously answer with handheld devices. The teacher displays distributions of answers and interesting particular answers (making sure that anonymity of the students is intact). The students have the opportunity of observing what their peers say within a minute. The teacher models good information retrieval strategies and evaluates the quality of the information presented on the web.

(4) **Internet problem.** A manager in a corporation has difficulties installing a new version of an internet browser and needs help. Instead of calling the software company, the manager learns about the browser with a QG&A facility that allows mixed initiative dialog with speech-to-text recognition and text-to-speech production. Some basic facts about the internet and browser are weaved into the conversation, but the most pressing goal is for the manager to solve the problem of installing the new browser by implementing the correct procedure that is somewhat tailored to the idiosyncrasies of the manager's system.

(5) **Collaborative composing of multiple choice questions.** Army personnel need to be trained on recent technologies released in the army. They normally would learn the material by reading uninspiring technical manuals. However, a TEAMThink project (Belanich, Wisher, & Orvis, 2003) has groups of 2-4 Army trainees learn from text by collaboratively composing multiple choice questions on the material. They critique each other's questions and response alternatives. Each group scores points by generating good questions, as evaluated by an expert or by the discriminative validity of the questions. The impact of this collaborative question generation technique is assessed with respect to learning gains and learner satisfaction (compared with solitary reading).

The above five scenarios are representative of the sort of learning environments that would benefit from QG&A facilities. The goals, content, depth, and questions are rather different among the scenarios. As a consequence, a different repertoire of technologies and tools would be needed to support question-based learning in these different learning scenarios.

Given that it is a mistake to treat all knowledge representations, cognitive processes, and learning scenarios as equivalent, there needs to be some principled or functional scheme for classifying the knowledge, skills, scenarios, and/or material to be learned. An evaluation of QG&A tools is contingent on which of the classes are being targeted, however these classes are eventually defined by the research community. Should the research roadmap accommodate a few, many, most, or all the classes that end

up being declared? Should some classes be explored earlier than others? Answers to these logistical questions are needed early in our planning.

4.3 Classification of questions

Schemes for classifying questions and analyzing the qualitative characteristics of questions have been proposed by researchers in education (Ciardiello, 1998; Dillon, 1984; Flammer, 1981), psychology (Graesser, Person, & Huber, 1992; Graesser & Person, 1994), and computational linguistics (Lehnert, 1978; Schank, 1986; Voorhees, 2001; Webber, 1988). One of the question taxonomies commonly used in the literature is Lehnert's (1978), later further developed for the educational field by Graesser and Person (1994). The Graesser-Person taxonomy is both grounded theoretically in cognitive science and has been successfully applied to a large number of question corpora in learning environments (such as human and computer tutoring, questions asked while using a new computer system, questions asked while comprehending text) and other discourse settings (questions raised in television news, questions by letters to an editor, questions in business transactions). Table 3 presents the 18 questions categories that are based on question content and that are sincere information-seeking (SIS) questions.

Table 3: Question taxonomy proposed by Graesser and Person (1994).

QUESTION CATEGORY	GENERIC QUESTION FRAMES AND EXAMPLES
1. Verification	Is X true or false? Did an event occur? Does a state exist?
2. Disjunctive	Is X, Y, or Z the case?
3. Concept completion	Who? What? When? Where?
4. Feature specification	What qualitative properties does entity X have?
5. Quantification	What is the value of a quantitative variable? How much? How many?
6. Definition questions	What does X mean?
7. Example questions	What is an example or instance of a category?.
8. Comparison	How is X similar to Y? How is X different from Y?
9. Interpretation	What concept or claim can be inferred from a static or active pattern of data?
10. Causal antecedent	What state or event causally led to an event or state? Why did an event occur? Why does a state exist? How did an event occur? How did a state come to exist?
11. Causal consequence	What are the consequences of an event or state? What if X occurred? What if X did not occur?
12. Goal orientation	What are the motives or goals behind an agent's action? Why did an agent do some action?
13. Instrumental/procedural	What plan or instrument allows an agent to accomplish a goal? What should an agent do next? How did agent do some action?
14. Enablement	What object or resource allows an agent to accomplish a

	goal?
15. Expectation	Why did some expected event <u>not</u> occur? Why does some expected state <u>not</u> exist?
16. Judgmental	What value does the answerer place on an idea or advice? What do you think of X? How would you rate X?
17. Assertion	A declarative statement that indicates the speaker does not understand an idea.
18. Request/Directive	The questioner wants the listener to perform some action.

The categories in Table 3 are defined according to the content of the information sought rather than on question signal words (*who, what, why, how, etc.*). The question categories can be recognized by particular generic question frames (which are comparatively distinctive), but not simply by signal words (which are ambiguous). Table 3 provides a question category label and the generic question frame for each category. Important to note is the distinction that the taxonomy makes between shallow and deep comprehension questions. Categories 1-8 are shallow comprehension questions that do not require a deep insight into the topic. Categories 9-16 on the other hand are deep comprehension questions that require more than dictionary or encyclopedic knowledge. In addition, inferences are needed to answer the deeper questions. It is these deep comprehension questions that help learners in constructing knowledge that supports the deeper levels of Bloom's taxonomy (levels 4-7 in Table 2). A computer system has been developed by Louwse, Graesser, Olney & TRG (2002) that classifies questions into these categories with up to 86% accuracy compared with human expert judges; this accuracy level is on par with human judges. The automated question classifier can quickly provide us with information about what kind of questions students ask. For example, in human-human tutoring dialog, the vast majority of questions (approximately 90%) are surface comprehension questions rather than deep comprehension questions. In other words, not only do students rarely ask questions, but they also ask the wrong type of questions. As we will argue later, one important milestone for any QA&G project in learning environments is to train students how to ask the right type of questions.

The Graesser-Person classification scheme is a start, but not the final word in classifying questions. One item on the QG&A research agenda is to investigate question classification schemes that satisfy specific empirical and computational criteria. For example, a classification scheme satisfies the empirical criterion of interjudge agreement to the extent that human experts agree on how to assign questions to particular question categories. A classification schemes satisfies most metrics of performance in computational linguistics if the algorithms can discriminate between categories and if the assignment of questions to categories show high recall and precisions scores (Voorhees, 2001), as defined shortly. It is anticipated that there is an optimal grain size in the question classification schemes, such that performance measures decrease when there are too few categories or too many categories.

It will be worthwhile to define a landscape of questions by considering the distribution of questions that occur in different categories of knowledge, cognitive

processes, and scenarios. If there are Q question categories, K categories of knowledge, P cognitive processes, and S scenarios, there are $Q \times K \times P \times S$ cells in the total space of questions. However, the questions in some cells are more appropriate, deeper, or sophisticated than others according to theories of learning, cognition, and inquiry. For example, causal antecedent and causal consequence questions are particularly appropriate for a physics tutoring scenario whereas instrumental-procedural questions are particularly appropriate for an Internet problem scenario; virtually any class of question would be needed for an electronic textbook on general science. Some theories of learning predict that only a subset of the cells in the QKPS space are quality questions (however defined by the theory) and that some cells are prone to occur empirically in investigations of learner inquiry. One item on the QG&A agenda is to analyze the space of questions in this landscape. What cells appear when the computer system or students ask questions in the different learning environments? How many of the cells advocated by learning theories end up being represented when data are collected? Obviously there is something radically wrong with a learning environment if the questions that it asks or the student questions it accommodates is totally misaligned with the questions that are considered important by the underlying pedagogical theory.

4.4 Measuring Progress

In order to track the progress in meeting the QG&A goals, there needs to be measures of performance and measures of the incidence of particular classes of observations. These measures require operational definitions that are valid, reliable, fair, and easy to compute. For example, the performance of a question answering facility is measured by some metric of the accuracy in which the system delivers high quality answers. The incidence of student questions is measured by the number of questions that a student asks per hour. There should be measurable improvement in relevant metrics in order to test claims that the QG&A facilities are improving over the course of research and development. This section identifies relevant metrics.

In the field of computational linguistics, state-of-the-art efforts have been evaluated under MUC (Message Understanding Conference), TREC (Text REtrieval Conference), and other large-scale research initiatives. Since 1992, TREC has supported research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC (Voorhees, 2001; <http://trec.nist.gov/>) is co-sponsored by the National Institute of Standards and Technology (NIST) and the Defense Advanced Research Projects Agency (DARPA) and contains various focus groups. One of these is the special track on question answering systems started in 1999 (<http://trec.nist.gov/data/qa.html>). A number of question answering systems are developed under the AQUAINT (Advanced Question and Answering for Intelligence) Program at ARDA (the Advanced Research & Development Activity) (<http://www.ic-arda.org/InfoExploit/aquaint/>). AQUAINT supports innovative, high-risk research to achieve significant advancements in technologies and methods for (a) understanding and interpreting complex questions, (b) determining the answers to

those questions, and (c) formulating and presenting the answers. AQUAINT is currently funding 16 separate research efforts, with plans to increase to 23.

The ARDA Technical Director (Dr. John Prange) has agreed that the AQUAINT and QG&A efforts can be supportive of each other. Our efforts will benefit from the AQUAINT's on-going research and its established relationships with researchers working in this area. AQUAINT is interested in broadening the impact of the AQUAINT R&D and will benefit from the insights and experiences of the learning technology community. The joint efforts will raise awareness of relevant tools and research results developed in areas other than education technology.

Measures of performance differ substantially among disciplines. The field of computational linguistics routinely collects recall scores, precision scores and F-measures (defined below) when comparing output of the computer with judgments of experts (Jurafsky & Martin, 2000; Voorhees, 2001). In contrast, cognitive scientists routinely apply signal detection theory and collect hit rates, false alarm rates, and d' discrimination scores (Green & Swets, 1966). Researchers in education and the social sciences report effect sizes and explained variance when measuring learning gains. All of these fields have some measure of interjudge reliability (such as Cohen's Kappa and Cronbach's alpha) when assessing the extent to which judges agree in making decisions. Every effort should be made to report measures that satisfy all of these fields because the QG&A agenda is inherently multidisciplinary. Some discussions and debates over appropriate computations of these metrics are healthy, but they should not distract the field from meeting its primary goals. One of the technical issues that needs to be settled early is to declare the metrics for evaluating progress on various research questions. The proposed metrics are defined below.

(1) **Incidence score**. Generally, the raw frequency of observations is adjusted for the document size. These incidence scores are the number of observations in category C per time frame t or per n number of words. For example, the incidence score for student questions would be an informative measure of whether a learning environment stimulates questions. These are sometimes called rate measures.

(2) **Interjudge agreement scores (also called interrater reliability)**. These metrics vary from -1 to +1, with +1 being perfect. For categorical scales, it is appropriate to report Cohen's Kappa, the agreement metrics most commonly used in computational linguistics. It controls for guessing at the base rate likelihood that a particular category occurs. For example, suppose that there is an agreement score of .90. An agreement score is the proportion of observations in which two judges assign the same category. A .90 would seem to be impressive, but it is not at all impressive when the proportion of observations in category A is .95 and the proportion in category B is .05; the agreement would be $(.95^2 + .05^2) = .905$ if the judges randomly guessed at the base rate likelihood. Cohen's Kappa scores adjust for the base rate. For continuous scales, it is appropriate to report Cronbach's alpha, which again varies from -1 to +1. Cronbach's alpha is a model of internal consistency, based on the average inter-item correlation. Under some

circumstances, researchers can report correlation coefficients for interjudge agreement, such as Pearson r or Spearman ρ .

(3) **Recall score.** Human experts normally serve as the gold standard when measuring the performance of computers in making decisions or producing output in category X. This measure is the proportion of observations in which the computer generates X, given that the expert humans generate X. Hence, the recall score = $p(X_{\text{computer}} | X_{\text{expert}})$. More general, recall is the proportion of correct answers given by the system divided by the total number of possible answers.

(4) **Precision scores (also called accuracy scores).** This measure is the proportion of observations in which the expert humans generate X, given that the computer generates X. Hence, the precision score is $p(X_{\text{expert}} | X_{\text{computer}})$. More general, precision is the total number of correct answers given by the system divided by the number of answers given by the system.

In addition to precision and recall a fallout measure is sometimes used that determines how spurious the system is. The fallout score is the total number of incorrect answers given by the system divided by the amount of spurious information in the text. Because fallout is far less common than the other measures, we will not include it in the list of metrics.

(5) **F-measure.** The F-measure is a combined overall score that takes both recall and precision into account. When recall and precision is weighted equally, it is computed as: $(2 \times \text{Recall} \times \text{Precision}) / (\text{Recall} + \text{Precision})$. However, it is also possible to weight recall and precision scores differentially. The F-measure varies from 0 to 1.

(6) **Hit rate.** This performance measure from signal detection theory is equivalent to the recall score in computational linguistics evaluations. The **miss rate** is defined as $(1 - \text{hit rate})$ because $(\text{hit rate} + \text{miss rate}) = 1$. The hit rate varies from 0 to 1.

(7) **False Alarm (FA) rate.** This is the proportion of observations in which the computer generates X, given that the expert human does not generate X. Hence, the false alarm rate = $p(X_{\text{computer}} | \text{not } X_{\text{expert}})$. The **correct rejection (CR)** rate is defined as $(1 - \text{FA})$ because $(\text{FA} + \text{CR}) = 1$. The FA rate varies from 0 to 1.

(8) **d' score.** This is a measure of discrimination that is highest when the hit rate is 1 and the false alarm rate is 0. It is a measure in standard deviation units between the distribution of scores in which the expert human says observation X occurs and the distribution of scores in which the human expert claims observation X does not occur. In practice, the d' score rarely exceeds 4, and is 0 when the computer is entirely unable to discriminate whether or not X occurs (compared with a human expert as a gold standard). There are alternative measures of discrimination, such as A_g scores, which vary from .5 (chance) to 1.0.

(9) **Effect size.** The effect size is particularly relevant to measurements of learning gains when researchers in education test a new experimental method of training. An experimental treatment (condition, group of subjects) is compared with a control condition when effect sizes are computed. The experimental treatment has a mean (M) and standard deviation (SD), and so does the control condition. The effect size is normally computed as $[(M_{\text{experimental}} - M_{\text{control}}) / SD_{\text{control}}]$. In essence, it is the improvement in scores when the experimental treatment is compared to the control, in standard deviation units. The selection of an appropriate control is normally a matter of considerable discussion and debate.

(10) **Variance of criterion measure explained by predictor measure.** The criterion measure (such as test scores) has a variance that is a product of multiple factors, including individual differences. The variance ranges from 0 to 1. This variance in the scores of the criterion measure C is explained by (i.e., predicted by, correlated with) some predictor variable of interest P . If the correlation between C and P is r , then the variance of C predicted by P is r^2 . The variance explained by P approaches 1 if it perfectly predicts C . R^2 is the index of variance explained in simple and multiple regression analyses, whereas *eta* is the index in analysis of variance designs with qualitatively different conditions.

(11) **Response time.** This is the amount of time it takes for the system to answer a question, present a question, or respond in some other fashion.

(12) **Confusion in learners.** Although this measure is the odd one out, because it is not a standardized metrics in the literature, it is equally important. There are a variety of indexes of learner confusion when learners interact with a learning environment. These include long pauses, counter-clarification questions, repeated actions that fail to advance the exchange, and so on. These indicators of confusion constitute another performance measure that needs to be collected in evaluations.

The above measures of performance do not exhaust the metrics and statistical indexes that can be used to track progress in the QG&A projects. However, they are important ones that need to be considered in evaluations. These metrics can be used to mark progress in evaluations of different question categories, learning scenarios, knowledge representations, and cognitive processes. Any particular project evaluation can focus on a subset of these measures.

4.5 Major design objectives of tools

There are major design objectives of the tools that will be developed in the QG&A research agenda. These include:

- 1) The learning environment that stimulates learner questions.
- 2) An interface for learners to ask questions.

- 3) Computational mechanisms for interpreting learner questions.
- 4) Computational mechanisms for answering learner questions.
- 5) An interface for the learning environment to answer learner questions.
- 6) Computational mechanisms for generating questions.
- 7) A facility for user modeling in interpreting and answering learner questions, as well as asking questions.
- 8) An interface for asking the learner questions.
- 9) An interface for learners to answer questions, both individually and collaboratively.
- 10) Computational mechanisms for interpreting learner's answers.
- 11) Linking people into the QG&A facilities.
- 12) Incorporating good pedagogy.

The remaining sections cluster these design objectives into groups and provide a research roadmap for each group.

5 RESEARCH ROADMAP

The QG&A research roadmap is organized into eight key research topics that collectively should facilitate effective question generation and answering, as we described for our design goals in Section 3. Associated with each research topic will be an array of software tools and a set of tasks (and subtasks) that need to be accomplished. The milestones for each task are projected for the next 3 years, 5 years, versus 10 years. In order to track progress in achieving the various tasks, there are recommended measures that include those defined in section 3.4.

5.1 Learning Environments that Stimulate Learner Questions, and Questions that Stimulate Learning

It is well documented that most learning environments do not stimulate many learner questions. According to the estimates of Graesser and Person (1994), a typical student asks .17 question per hour in a conventional classroom and 27 questions per hour in one-on-one human tutoring. They estimate that the upper bound in the incidence of student questions per hour is approximately 125, as in the case of the *Point & Query* software where the only way students can learn the material is to ask questions and comprehend answers to their questions (as described later). There obviously is a long way to go in identifying learning environments that encourage more inquiry and curiosity in the learner.

Available research has identified a number of characteristics of learning environments that stimulate curiosity and sincere information seeking questions, as opposed to questions that merely garner attention, monitor conversation flow, or serve auxiliary social functions. We know, for example, that the following characteristics of learning environment stimulate students' asking of sincere information-seeking questions.

(1) Learning environments that place the student in cognitive disequilibrium, as in the case of challenges of entrenched beliefs, obstacles to goals, contradictions, anomalous events, breakdown scenarios, salient contrasts, decisions in the face of equally attractive options, game theoretical dilemmas, and salient deviations from norms (Chinn & Brewer, 1993; Dillon, 1988; Graesser & McMahan, 1993; Graesser & Olde, in press). Computational models have been proposed which specify what questions are asked when confronted with different forms of cognitive disequilibrium, as in the case ABE (Kass, 1992), SWALE (Schank, 1986, 1999) and PREG (Otero & Graesser, 2001).

(2) Learning models that didactically train or model the asking of questions (King, 1989, 1994). Modeling good question asking can be accomplished by expert human models or peers (Palincsar & Brown, 1984), animated conversational agents on computers (Craig, Gholson, & Driscoll, 2002), or the set of questions presented on a question menu (Graesser, Langston, & Baggett, 1993).

(3) Inquiry based learning environments that encourage hypothesis testing and experimentation for achieving long-term objectives on authentic problems (Linn & Hsi, 2000; White & Frederiksen, 1998).

Although there is evidence that the above learning environments increase the incidence of student questions, there needs to be more systematic research on the particular characteristics of learning environments that trigger particular categories of questions. Therefore, the first task is to document the relationship between features of different learning environments [signified as $F(LE)$] and the landscape of questions (signified as QKPS, as defined in section 3.3). This mapping [$F(LE) \rightarrow QKPS$] can be quantified empirically by computing question incidence scores for particular cells in the landscape of questions. The proposed tasks will document these mappings for a representative set of learning scenarios in science, mathematics, engineering, and technology (see section 3.2 for examples):

Proposed classes of learning environments:

- (1) reading from texts (for ultimate use in an electronic textbook)
- (2) intelligent tutoring systems
- (3) classroom interactions
- (4) solving technological problems at work
- (5) collaborative learning or collaborative work.

The subject matter domains will not be highly constrained during the first three years because the goal is to explore a broad diversity of learning situations, as long as the above 5 learning environment categories are represented. Subject matters will be selected more strategically at the 5-year and 10-year points to the extent that the $F(LE) \rightarrow QKPS$ mapping is better understood both theoretically and empirically. It should be noted that there is a subtle difference between a learning environment and a learning scenario. A learning environment has generic features and capabilities whereas a learning scenario is more concrete, context-constrained, and situated in practice.

Theories of question asking and pedagogy can be tested by evaluating how discriminating the theories are in predicting what cells in the landscape of student questions or system questions are filled with above-zero incidence scores. We believe that this is a crucial step in building a more rigorous science of question-based learning. Any theory with teeth will offer discriminating predictions as to which of the cells should have above-zero incidence scores (or have high, medium versus low incidence scores if researchers desire some continuum of values). The success of the theory can be rigorously quantified by computing recall, precision, F-measures, hit rates, false alarm rates, and d' scores that compare predicted cell values with empirical incidence scores. Theories can be compared and tracked on these metrics over time as the research agenda evolves. In summary, therefore, there will be tasks for testing theories of question asking and pedagogy. The most rigorous tests will be during the 5-year and 10-year points.

Questions stimulate learning. There is ample evidence that comprehension and learning improves after students are trained how to ask question particular types of questions (King, 1989; 1992, 1994; Rosenshine et al., 1996). However, there needs to be more systematic research on what knowledge (K) and cognitive processes (P) are improved by training of particular classes of questions (Q) in particular classes of learning scenarios (S). For example, training learners how to ask causal antecedent (*why*) and causal consequence (*what-if*) questions presumably promotes deep learning of causal structures when learning physics in a tutoring system, whereas instrumental/procedural (*how*) questions promote shallow learning of goal-oriented procedures. Such claims can be tested quantitatively by conducting controlled experiments that compute effect sizes of manipulations (i.e., comparing experimental and control conditions) on dependent measures (called outcome measures). The alternative approach is to conduct correlational naturalistic studies that compute the amount of learning variance that is explained by $SQ \rightarrow KP$ relationships. Therefore, there will be tasks that investigate the relationship between the training of question asking skills and the learning of the material.

Table 4 presents the tasks that can be defined for learning environments that stimulate learner questions, and questions that stimulate learning. As the table shows, the tasks are comparatively easy for most of the milestones in the first 3-years because less is known or basic knowledge needs to accumulate. As the years advance, the tasks assigned become progressively more difficult. It would be worthwhile to start out a broad array of learning environments with respect to complexity, the depth of knowledge required, and social organization. This is necessary to arrive at a landscape of questions that is both representative and broad in scope. However, the challenges and required sophistication of the tools are expected to increase over the years. Nevertheless, some of the tasks will never end up being complex because of inherent technical difficulties and scientific challenges. As pointed out before, performance of each of these tasks and milestones will have to be carefully measured. The last column in Table 4 indicates the measures of performance to be used, as described in section 3.4.

Bringing industry and academia together by working towards more unified goals would be a general milestone to achieve early in the research agenda. Research centers at universities, industry and the military have developed learning environments that afford question asking and inquiry. Large-scale projects on educational technology have been funded by NSF, ONR, ARI, McDonnell Foundation, and Spencer Foundation on grants at approximately a dozen universities. Major efforts on building advanced learning environments have been supported at a small number of large corporations (Apple, Hewlett Packard, IBM, Microsoft). Despite the large number of projects, the technologies and information base are very distributed.

Tasks	Milestones			Measures
	3-year	5-year	10-year	
Investigate $F(LE)$ → QKPS mappings (which features of learning environments elicit questions)	Run 2-3 studies of how varying features in a learning environment changes quantity and types of questions Documentation of critical components of learning environments that stimulate question asking from the studies	Identification of 10 specific subject matters for the learning environments and studies run on those subject matters with specific learning scenarios Eye-tracking and other fine grained measures in use Focus on specific components of learning environments that stimulate question asking from the studies	100 studies; assessing types of questions elicited by a learning environment becomes a routinely used dependent variable	1, 2, 3, 4, 5, 6, 7, 8
Test and compare theories that predict landscape of questions in learning environments and scenarios	Integration of existing learning theories and the development of refined theories (pedagogical, cognitive, discourse) for landscape of questions Initial testing of theories in the most promising cells of the landscape of questions	Expansion of theories to include impact of training to ask questions and personality factors Theories compared to each other and refined	Fairly robust theories and models exist providing accurate (within 10%) predictions across majority of cells of the landscape of questions	1, 2, 3, 4, 5, 6, 7, 8
Investigate learning from questions, including SQ → KP mappings	2-3 studies of learning from questions in controlled and naturalistic environments (same studies as above to measure learning gains) Development of initial theories of how knowledge and cognitive processes are affected by scenarios and training	More careful analysis of question content and the content of the gains is used to refute earlier theories of what aspects of question generation cause learning. 10 studies Theory and model development of how knowledge and cognitive processes are affected by scenarios and training	Fairly robust models developed on the effects of questions on cognitive processes Models provide fairly accurate predictions of learning based on the effects of scenarios and prior knowledge (within 20%)	1, 2, 9, 10
Develop and refine question taxonomy	Existing question taxonomies revised and refined Domain specific aspects better understood Investigation of additional learning environments to ensure complete coverage.	Question taxonomy linked to example questions from question corpus Investigation of additional learning environments to ensure complete coverage	Minor revisions to taxonomy made based on novel learning environments	2; use in the literature
Document landscape of questions (QKPS) in 5 classes of learning environments and scenarios	Initial repository of example questions covering landscape (both questions asked and answered) developed and linked to question taxonomy Examination and refinement of the distribution of questions across the QKPS and identification of most promising cells in the distribution for research. Development of tools to help annotators of corpus. (similar to efforts in table 8)	Collection of corpora of example questions across taxonomy for all 5 classes of learning environments is completed Question corpus annotated by hand for linguistic, knowledge, pedagogical information. (similar to efforts in table 8) Development of techniques for automated annotation of corpora can annotate within 10% accuracy of human coding. similar to efforts in table 8)	Complete repositories of example questions across taxonomy for additional new classes of learning environments Completion of techniques for automated annotation of corpora can annotate as accurately as human coders	1, 2

The Learning Federation www.thelearningfederation.org

For more information contact: Kay Howell (khowell@fas.org)

Table 4: Stimulating Learner Questions and Questions that Stimulate Learning

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure

5.2 Interface for Learners to Ask Questions

The human-computer interface needs to be designed to make it easy for students to ask questions and to provide some guidance on what sort of questions can be (or should be) asked. The goal is to have a very small barrier between the moment that a learner has inquiry (e.g., curiosity, a hypothesis, cognitive disequilibrium) and the moment that a question gets entered into the learning system. This barrier is reduced to the extent that: (a) there is a short time duration, (b) the system can accommodate a large space of questions that learners are likely to ask, (c) the learner knows what categories of questions the system can handle, (d) it is easy for the learner to learn how to input questions, (e) good relevant answers are delivered quickly, (f) the learner understands the answers, and (g) the learner can quickly modify questions. For the first time in history, technology is available that can radically reduce these barriers to the point where most questions can be asked and answers received within a few seconds. It is analogous to a ubiquitous private tutor being available for each student.

A large number of interfaces have been designed for question asking. Each interface has strengths and limitations, as enumerated below. They are described in the order of sophistication in terms of fewer user constraints. It is important to acknowledge that many of these question facilities also incorporate components of question interpretation, question answering, and various facilities that go beyond the learner question interface per se. The development and evaluation of these other computational modules are addressed in subsequent subsections.

(1) **Structured Query Languages (SQL)**. There is a constrained syntax for entering questions, usually in the form of a Boolean (*and*, *or*, *not*) combination of well-formed expressions with functions, variables and values. SQL is desired when there is a need to precisely specify the questions, and the expected answers are selected from highly structured databases. However, it takes a long time to learn SQL and most learners are unwilling to invest the time. It also takes a long time to enter a particular query. SQL is useless when the database is unstructured. Most individuals do not have the aptitude for analytical reasoning that SQL requires.

(2) **Structured queries guided by menus, GUI's, or templates**. There is a front-end that guides the user in constructing queries that get transformed into SQL form. Examples of these facilities are query-by-example and retrieval by reformulation. This approach is much better than pure SQL for beginning and infrequent users of a system. It helps the users in entering the query by providing them a little bit more flexibility.

(3) **Frequently Asked Questions (FAQs)**. The learner selects a question from a list of questions supplied by the system. A canned answer is immediately presented after a selection is made. FAQs are perhaps the most common interface for asking questions and are useful when the knowledge domain is limited and the list of questions covers most of the questions that users would ever ask. Unfortunately, it is too often the case that designers of the systems create an inventory of questions that users rarely ask, and

miss the questions that users frequently ask. Furthermore, users have to evaluate the existing FAQs in order to select the question that best suits their needs. A good FAQ would dynamically change the menu selection over time by somehow collecting data on what questions users really ask and by having experts answer these questions (normally asynchronously). Examples of the dynamic FAQ's are Ackerman's Answer Garden, various organizational memory systems, and web forums. Ideally, such a self-organizing FAQ facility would eventually cover over 90% of the questions that users would ever ask. One advantage of FAQs is that users can learn what questions are important.

(4) **Point & Query (P&Q)**. The user points to (or clicks on) a hot spot on the display and then a menu of relevant questions immediately appear. Then the user selects a relevant question and the answer immediately appears. The questions are tailored to the idiosyncratic content of each display so there is a context-sensitive menu of questions. Questions can range from being very specific (*What causes friction?*) to being generic frames (*What causes X?*). P&Q interfaces (Graesser et al., 1993; Graesser, Hu, Person, Jackson, & Toth, 2003; Schank, Ferguson, Birnbaum, Barger, & Greising, 1991) can present both frequently asked questions and the ideal questions that an expert hopes the user will ask. Therefore, P&Q helps the user learn what good questions are under particular contexts, an indirect form of training and modeling of good question asking skills. Questions are asked moderately quickly by just two clicks of a mouse. P&Q, however, suffers from similar weaknesses as the FAQs. It breaks down when there is a mismatch between available questions on the system and actual questions that users ask. The questions and answers also need to be handcrafted by the designer, in a structured database.

(5) **Open-ended questions on search engines**. Good examples of these search engines are Google and Yahoo. The user types in a string of words that need not be well formed. Relevant web sites are quickly accessed in a priority order. These systems have provided an excellent solution to one major hurdle in question answering: Fetching relevant text documents. Also, the questions can be posed in natural language rather than well-formed formal expressions from which the system selects the relevant keywords (e.g. ignoring function and other frequent words). However, there are also limitations with these search engines. The learner is not exposed to example good questions and to strategies of modifying their queries to get better answers, so the user does not learn good question asking skills. Search engines miss a large percentage of relevant documents and present a large percentage of irrelevant documents. The engines return texts but do not extract information from the retrieved documents. So users do not really get an answer to the question (i.e., extracted from the document), but instead receive the full document that hopefully contains an answer to the question. Finding the right document and the requested information in that document can be painstaking.

(6) **Natural language questions with keyboard input**. Unlike the previous interfaces, natural language questions allow for substantially more user flexibility. Users can ask questions in their own words and are not limited to particular question formats. Instead, a question is parsed and semantically interpreted by modules in computational linguistics and artificial intelligence (Allen, 1995; Jurafsky and Martin, 2000) and an

answer is returned in some systems. The performance of these natural language query systems has reached the point when they can be incorporated in learning environments, although performance is not as impressive for deep questions as for the shallow questions. However, learners will be frustrated if their questions are not interpreted correctly and if incorrect answers are produced; expectations are also higher than for FAQs. It is important for the system to allow the user to verify that questions are received correctly (e.g., *Is your question...?*), but this has the unwanted side effect of adding time and tedium. Despite the recent advances in computational linguistics, question answering systems are still prone to errors, for instance in ambiguous questions (e.g. what are the most recent children books? (books by/for/about kids?) or questions in which common ground and implicatures are assumed (how are you today?).

(7) **Natural language questions with speech recognition**. Text-to-speech facilities have become sufficiently impressive to try to integrate with learning environments. There are now a number of systems that accommodate continuous speech in conversational dialog and perform rather well, such as those developed at University of Colorado (Pellom, Ward, & Pradhan, 2002), AT&T (Litman, Walker, & Kearns, 1999), and Carnegie Mellon University (Lamere, et al., 2003). However, these systems perform best in very restricted conceptual domains, such as planning flight schedules and or helping callers find telephone numbers. It is an open question how well they will perform with more complex learning environments; that would need to be evaluated in the proposed QG&A roadmap. Furthermore, they suffer from many of the same problems as the interfaces with keyboard input.

(8) **Multimedia input**. Questions may be posed by a combination of speech, gestures, pointing, and actions. For example, one could imagine manipulating controls on a simulation and speaking simultaneously, e.g., *What happens when I change this?* while moving the control. One can point on a touch panel screen or draw a path on the screen with a stylus, while saying *Is this how I should drive through the city?* Multimedia question input is only beginning to be investigated, but it is an important class of direct manipulation interfaces to explore.

(9) **Questions implicit in actions**. Questions can sometimes be inferred from the actions that a learner performs in a simulation environment. Suppose that a learner repeatedly increases the value on a simulation control and then observes a component this is displayed in an animated simulation of a device. The pattern of action and eye tracking suggests the question *How does control X affect the component Y?* An answer to this question could be generated by the system without the learner ever having to express the question in language. As with the multimedia input, this interface is only in its initial stages, but projects like the BlueEyes system (MIT, University of Memphis) that measures facial actions and emotion are initial steps toward questions implicit in actions.

There are tradeoffs in using each interface because of their associated strengths and limitations. For example, Table 5 summarizes how the nine interfaces, as they have been developed so far, compare on five performance criteria: Accuracy of question interpretation, speed of asking question, training time, facilities for improving questions (e.g., giving example good questions), and amount of material that is covered. The values of these cells are estimates by some members of this Roadmap.

Table 5: Current Achievement of Performance Criteria of 9 Classes of Interfaces

Interface	Accuracy	Speed	Training	Improving Questions	Material Covered
Structured Query Languages (SQL)	Moderate	Slow	Slow	Low	Moderate
SQL with menus	High	Slow	Moderate	Moderate	Moderate
Frequently Asked Questions	High	Moderate	Fast	High	Small
Point&Query	High	Fast	Fast	High	Small
Search engines	Low-Moderate	Moderate	Fast	Low	Large
Natural language questions– keyboard	Low	Moderate	Fast	Low	Moderate
Natural language questions – speech	Low	Fast	Moderate	Low	Small
Multimedia input	Low	Fast	Fast	Moderate	Small
Questions implicit in actions	?	?	?	?	?

The primary research goals are to develop systems in which (a) learners can quickly ask questions and (b) learners can learn how to improve their questions. Table 6 summarizes the tasks for question interfaces.

The research agenda would be to design, develop and evaluate these learner question interfaces in the classes of learning environments that were illustrated in Section 3.2: An electronic textbook, computer simulation with hypothetical reasoning, classroom analyses of problems, on-the-job solutions to technical problems, and collaborative composition of questions. In order to meaningfully compare the interfaces, one would need to control the subject matter content. So the best studies might involve evaluating 2 or more interfaces on the same learning content and also comparing 2 or more different classes of learning environments. The choice of learning environments and learning contexts is nontrivial. For instance, for highly structured data SQL languages might be preferred over natural language questions with speech recognition, whereas such a method should be avoided in free text. For some of the interfaces, the content and type of content might preclude certain interface choices. The question of what learning environments and learning content to select in the research roadmap will therefore be discussed and hopefully resolved at the next workshop.

The development of authoring tools is needed for expanding the repertoire of questions and answers. The search engines and many natural language query systems do not require special authoring tools, unless some form of mark-up language is needed for the documents in the document space. In contrast, the FAQs and P&Q require authoring tools for handcrafting questions and answers in the proper format the system can handle. The incidence and speed of adding new questions would be the relevant data to collect in evaluations of the authoring tools.

Table 6: Interfaces for Learner Questions

Tasks	Milestones			Measure s
	3-year	5-year	10-year	
Maximizing the speed of asking questions	<p>Student can type questions and describe the context (e.g., paragraph, or graphic in text)</p> <p>System echoes its interpretation of the question back to the user for verification</p> <p>Spoken language questions in limited domains.</p>	<p>Student can type questions and enter a description of where they are in solving an exercise in the text</p> <p>Intuitive interface allows user to correct the system's understanding of the question (e.g., correct word sense error or understanding of ambiguous phrases)</p> <p>Spoken language questions in focused domains.</p>	<p>Student can scan relevant part of text and transmit spoken questions</p> <p>Intelligent system learns from previous interaction with users and improves its question understanding</p> <p>Spoken language questions in open domains.</p>	1,11,12
Maximizing the coverage of the landscape of questions (QKPS) in learning environments and scenarios	Utility for authored environments (text, ILE, classroom) that transmits all student questions to the author or publisher along with student report of adequacy of answer	Utility that enables users (students) to exchange questions and answers (and transmits exchanges to database maintainer)	Utilities for collaborative authoring of and inspection of shared databases.	1,12
Evaluating impact on learning	"Usability" studies of interface features examine question-asking frequency with and without the feature	Formative evaluations of interface features examine average question quality and changes in question quality (learning to ask good questions) with and without the feature	Summative evaluation of interface features that examine domain learning gains with and without the interface feature	9,10
Developing the documents and databases that provide questions and answers for the designated interface	<p>Establishing standardized Q&A content formats across all use environments</p> <p>Markup language for supporting question answering from large text collections.</p>	<p>Utility for all environments that enables authors to associate answers with questions and coaches authors on making answers maximally discriminable</p> <p>Large text collections manually annotated with QA markup language.</p>	<p>Utilities for annotating images for student use in deictic reference</p> <p>Automated markup of large text collections in support of question answering.</p>	2, 3,4,5,7,6
Maximizing ease of learning how to ask questions	<p>Scaffold question asking with mini video tutorials and FAQ models</p> <p>Student's natural language questions are evaluated and alternative questions are presented, based on predefined question formulae</p>	Learning environments coach students on ambiguities in question form, including referents	<p>In returning an answer set for a question, for each answer return a modification of the initial question that most strongly elicits the answer</p> <p>Student's natural language questions are evaluated and alternative questions are generated</p>	1,11
Helping the learner to ask better questions	Author provides database of FAQs that model good questions	Primary ILE tutor help consists of FAQs that model good questions from which the student can select	ILE Tutor employs model of student knowledge to coach student on questions that should be asked	1,2,9,10, 11,12

The Learning Federation www.thelearningfederation.org

For more information contact: Kay Howell (khowell@fas.org)

Authoring tools that integrate QA facilities with LE and scenarios	Student or tutor can return to any problem state in the solution process for post-problem review session	Student can mark the context and deictic references in the record (video, audio) of a classroom lecture	Student can build up a useable record of questions asked and answered	1,11,12
--	--	---	---	---------

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time

5.3 Interfaces for Asking Learners Questions and Answering Learner Questions

The computer system can output information in many forms, structures, media, and layouts. There are conventional static media of illustrated texts (print, pictures, figures, tables, and diagrams) and animated media (animated simulations of complex systems, video clips, audio recordings). Animated conversational agents produce language with synthesized speech (text-to-speech), facial expressions, pointing and gesture (Johnson, Rickel, & Lester, 2000). Complex multimedia consists of hybrids of the two or more forms of media, hopefully in a coordinated manner. For example, one could imagine a system that answers a question by showing an animated simulation of a device and an animated conversational agent that explains to the user what is happening as the device animation unfolds. The multimedia display would hopefully answer the question.

During the last decade, cognitive psychologists have investigated the impact of modalities and multimedia on learning and the learners' impression of the learning experience (Atkinson, 2002; Collins et al., 2000; Mayer, 1997, 2002). They have identified conditions in which it is appropriate to present information in single or multiple modalities (text, pictures, sound), to present information contiguously in time and space, and to avoid split attention effects. The QG&A agenda should investigate alternative multimedia design alternatives as they incisively pertain to computers asking and answering questions. For example, when is it best to deliver information in printed text versus speech (Olson, 2002; Whitaker, 2003), in language versus highlighted pictures (Schnotz, Bannert, & Seufert, 2002), or in static illustrations versus animated simulations (Hegarty, Narayanan, & Freitas, 2002; Pane, Corbett & John 1996). It is important to clarify the relevance of the multimedia alternatives to question asking and answering mechanisms.

There are different formats of the questions asked by computer systems. The designers of traditional computer-based training were fond of multiple choice questions, true-false questions, ratings, and short answer questions. The answers to these questions could be scored quickly, objectively, and mechanically. There has been a large enterprise of psychometrics and testing, which has guidelines on how to develop test questions and response alternatives that maximize the reliability and validity of the items. More recently, researchers have designed multiple choice question alternatives in a fashion that diagnoses users' misconceptions (Hestenes, Wells, & Swackhamer, 1992; Hunt & Minstrell, 1996). Thus, the selection of particular wrong answers is as informative as the likelihood of giving a correct answer. Graesser and Wisher (2002) review the alternative question formats and guidelines for developing the questions.

Whenever the interface has non-print media, there is the pressing question of whether the learner understands the function and meaning of the symbols, icons, GUI, or layout. The comprehensibility and usability of interfaces are systematically investigated in the fields of semiotics (Martins, 2002), human factors, and human-computer interaction (Card et al., 1983; Rosson & Carroll, 2002; Dix, Finlay, Abowd, & Beale, 2002; Norman & Draper, 1986; Schneiderman, 1987). Experts have developed guidelines for developing interfaces that promote usability and user satisfaction.

Researchers perform qualitative analyses by asking users to think aloud, ask questions, or answer questions while they perform benchmark tasks with the system. Comprehension difficulties and troublesome bottlenecks in the interface are manifested by analyzing these verbal protocols, by observing the users as they work on the benchmark tasks, and by collecting logs of the user's normal interaction with the system. Quantitative analyses are also conducted on errors, response times, and users' ratings of their impressions of the system. The methods developed in the field of human factors and human-computer interaction can be directly applied to evaluations of the QG&A interfaces.

Table 7 presents tasks for building interfaces that have the computer ask questions and answer learner questions. The final column specifies the measures whereas the remaining three columns designate the difficulty of level of the tasks at 3-, 5-, and 10-years.

Table 7: Tasks: Interfaces for Asking Learners Questions and Answering Learner Questions

Milestones				
Tasks	3-years	5-years	10-years	Measures
Maximizing the speed and intelligibility of different interfaces	Answers in text and speech.	In focused domains, answers are reworded or expanded if user indicates confusion or lack of understanding.	In broader domains, answers are reworded or expanded if user indicates confusion or lack of understanding.	1,2-8, 11, expert guidelines
Evaluating the impact of interfaces on learning	Currently available systems tested	Interfaces accommodate learning, incl. new technologies	Latest technologies implemented	9, 10
Evaluating the impact of interfaces on usability and learner impressions	“Usability” studies of interface features examine errors, response times, and users’ ratings of their impressions of the system with and without the feature	Formative evaluations of interface features examine average answer quality and learner satisfaction with and without the feature	Summative evaluation of interface features that examine domain learning gains with and without the interface feature	Qualitative measures from verbal protocols; methods from HCI and human factors
Building software utilities & authoring tools that integrate Q&A interfaces with LE and scenarios	Utilities and tools based on selection of existing theories (development drives selection of theory)	Utilities and tools based on selection of best learning theories (theory drives development)	Conformity to standards, implementation in learning systems using a variety of interfaces	1,11

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d’ score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time; 12=learner confusion

5.4 Comprehending and Answering Learner Questions

ARDA AQUAINT (Harabagiu, et al., 2002; Pasca & Harabagiu, 2001) and the TREC Question Answering Track (Voorhees, 2001; Voorhees & Tice, 2000) have reviewed the recent progress that has been made in building computer systems that attempt to comprehend questions, access relevant documents, and extract answers from the documents. These are extensions of some large-scale initiatives, funded by DoD, that have systematically measured the performance of different systems in their ability to access relevant documents from large document corpora (called information retrieval, IR) and to extract lexical, syntactic, and semantic information from text (called information extraction, IE). The performance of IR systems is assessed in TREC (Text REtrieval Conference). The performance of IE systems has been assessed in the Message Understanding Conferences (MUC, http://www.itl.nist.gov/iaui/894.02/related_projects/muc/), the DARPA Hub-4 Broadcast News evaluation, the 1999 DARPA-TIDES Information Extraction-Entity Recognition (IE-ER) evaluation, and the Automatic Content Extraction program (<http://www.nist.gov/speech/tests/ace/index.htm>). The National Institute of Standards and Technology (NIST) is a neutral party that selects the benchmark tasks, performance measures, and scheduling of the assessments. This assures that the performance of dozens of different systems can be evaluated and compared with fairness and objectivity. The different systems are compared quantitatively on various capabilities, although it is explicitly emphasized that the goal is to mark progress in the field as a whole rather than to hold competitions.

The distinctive focus of the ARDA AQUAINT and the TREC QA track is evaluation of question answering systems. A sample of questions is asked of each system and the answers are recorded. Typically, the top 5 answers are generated for each question, in a rank ordering. The analysts measure the likelihood that the correct answer is included among the top 5 answers or there is a weighting scheme that assigns more weight when the correct answer has a higher rank (i.e., 2nd as opposed to 5th). A corpus of documents is declared in each evaluation, such as articles on terrorism (a small specialized document space), the *Wall Street Journal*, or documents on the web (a wide open document space). The performance of the QA systems has often been extremely impressive, particularly for closed class questions (e.g., *who*, *what*, *when*, *where*) as opposed to open-class questions (e.g., *why*, *how*). A good example of one of these Q&A systems is the one developed by Harabagiu and colleagues (Pasca & Harabagiu, 2001; Harabagiu, Maiorano, & Pasca, 2002).

A recent QA roadmap (Burger et al., 2001) identified 12 areas in which research is required, along with a set of subtasks and 5-year milestones for each. These 12 areas are listed below:

- (1) Question taxonomies
- (2) Semantic models of question understanding and processing
- (3) Incorporating user context into question answering strategies

- (4) Heterogeneous data sources
- (5) Answer justification
- (6) Answer formulation (display)
- (7) Real-time question answering
- (8) Multilingual question answering
- (9) Interactive question answering
- (10) Advanced reasoning for question answering
- (11) User profiles for question answering
- (12) Collaborative question answering

There are systematic methods of tracking progress on each of these components, as documented in the QA Roadmap.

It would be prudent for this QG&A roadmap to build on the QA roadmap and adopt the same evaluation methods. Ideally, learning environments would be declared as one of the data sources in a future test of the question answering systems of ARDA AQUAINT. The groups that have participated in ARDA AQUAINT would presumably be encouraged to participate in the QG&A initiative. There would be collaborations between the designers and testers of question answering systems in ARDA AQUAINT and the researchers who focus on learning and questions in QG&A.

The QG&A agenda would require additional layers of research and evaluation beyond those of ARDA AQUAINT. In particular, there are three layers. First, it is important to maximize the coverage of learner questions that are asked or should be asked from the standpoint of pedagogy. Second, there needs to be an evaluation of the impact of the question comprehension and answering facilities on learning gains. Third, there needs to be utilities and authoring tools for marking up the corpus of documents that serve as information sources for the answers. Every attempt should be made to establish standards for representing, marking up, and organizing the content in the learning repository, following the missions and standards of ADL/SCORM (see also section 3.1) and the Open Language Archives Metadata Set (OLAMS). Table 8 presents the tasks on the roadmap for comprehending and answering learner questions.

Table 8: Comprehending and Answering Learner Questions

Tasks	Milestones			Measures
	3-year	5-year	10-year	
Answer formulation	Basic answer formulation for concept completion, feature specification, and quantification questions. Answers extracted from existing sources	Basic answer formulation from a single source for all categories of questions	Complex answers compiled, merged, and generated from multiple sources	1,2,3,4,5,6,7,8
Response time	Expert replies within one day (human or FAQ's) Electronic information within 60 seconds for all question categories	Expert replies within an hour Electronic information within 10 seconds for all question categories	Expert replies within 5 seconds Electronic information within 2 seconds for all question categories	11
Building software utilities and authoring tools for marking up documents in learning repository	Software tools need to optimize human-computer interface, with adequate help and training for <users> domain experts with minimal knowledge of AI, natural language, and cognitive science Users of authoring tool are capable of completing mark up's successfully.	Identify metadata that users can versus cannot construct while using the authoring tools Users of authoring tool are capable of completing mark up's successfully and quickly	Perform cost-benefit analyses of particular mark-up languages and interfaces Users of authoring tool are capable of completing mark up's successfully, quickly, and with high quality	1,11, conformity to ADL SCORM and OLAMS standards
Multilingual question answering	Question answering in languages other than English	Questions asked in English, answers located in text sources from another language and translated back to English	Questions asked in English, answers located in text and audio sources from another language and translated back to English.	1,2,3,4,5,6,7,8
Heterogeneous data sources	Answers from text	Answers from audio	Answers from video	

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time

5.5 Mechanisms for Interpreting Learner Answers

There are a variety of ways for learners to enter their answers when answering questions in a learning environment. These include clicking a mouse or pointing to an option on the display, typing in alpha-numeric answers, natural language expressions typed via keyboard, spoken natural language, entering input to handheld devices, and so on. It is easy to interpret learner answers that consist of mouse clicks on menu options and short sequences of alphanumeric characters entered by keyboard. It is substantially more difficult for the computer to interpret sentences, lengthy texts, visual sketches, and complex actions. However, information from all of these input modes eventually need to be interpreted, deeply comprehended, and evaluated pedagogically in any sophisticated learning environment. Fortunately, some progress has been made in interpreting the more complex forms of input even though the available systems are not perfect.

Extracting information from open-ended answers in natural language has not been automated until recently, with advances in computational linguistics (Baeza-Yates & Ribeiro-Neto, 1999; Jurafsky & Martin, 2000; Sparck-Jones & Willet, 1997) and evaluations in the Message Understanding Conference (MUC). Various models have been developed for information retrieval, including Boolean, vector and probabilistic models. In fact, the current systems developed in the ARDA AQUAINT Program and TREC show how successful and diverse these systems are. It would be prudent to piggy back on these advances in MUC. These researchers have developed and evaluated systems that assign syntactic classes to words (called part of speech tagging), that disambiguate words with multiple senses based on context, that assign referents to nouns and pronouns, that connect noun-phrases and prepositional phrases to verbs in appropriate case-structure roles, and that perform other forms of syntactic and semantic processing. These computational linguists have developed lexicons, ontologies, parsers, and semantic structures that support these computations. Therefore, the QG&A researchers need to team up with the community of computational linguists in TREC, MUC and AQUAINT.

One computational linguistic development that deserves special attention is Latent Semantic Analysis, because it has been proven to be useful in information retrieval as well as in knowledge representation and dialog maintenance. Latent semantic analysis (LSA) is a practical and valid statistical technique for representing world knowledge and performing some interpretive processes (Kintsch, 1998, 2001; Landauer, Foltz, & Laham, 1998). LSA was originally developed and tested in the TREC (Dumais, 1993), but more recently has been applied to discourse and natural language processing. The words, sentences, and paragraphs in a large corpus of texts (an encyclopedia or several books) are compressed into 100 to 500 functional dimensions. These dimensions are used for evaluating the conceptual similarity or relatedness between any two bags of words; a bag of words is one or more words that are not ordered in sequence or priority. LSA is a powerful statistical method for performing meaning-based pattern matching. When experts are asked to rate the relatedness of pairs of bags of words, their ratings are very close to LSA similarity measures (Graesser, Wiemer-Hastings, Wiemer-Hastings, Person

& Harter, 2000; Landauer et al., 1998; Olde, Franceschetti, Graesser, 2002). LSA has provided the foundation for grading the quality of essays and tracking the knowledge of learners. The Automated Essay Grader can grade essays as reliably as experts in composition (Foltz, Gilliam, & Kendall, 2000). Summary Street (E.Kintsch, et al., 2000) is an LSA-based system that assists children in writing summaries of articles by computing the coherence of the set of sentences and detecting gaps when compared with an ideal summary. AutoTutor (Graesser, Person, Harter, & TRG; Olde et al., 2002) evaluates the quality of student contributions during tutoring by matching learner verbalizations to expected good answers and misconceptions; AutoTutor's LSA component does this almost as well as graduate student research assistants. For the first time in history, it is conceivable to have computers accurately evaluate lengthy answers of students expressed in natural language. The accuracy will improve with advances in computational linguistics, with improved lexicons, syntactic parsers, reference resolution, semantic analysis, and discourse analysis.

It is much more difficult for computers to interpret complex nonverbal input. There are products that allow users to draw on a screen with a stylus, light pen, or finger. There are sensors that can accurately record the actions of people, their gestures, their facial expressions, and their eye movements. However, the fields of visual pattern recognition are very much at the infancy stage when it comes to accurately interpreting the visual input. Progress on this front is likely to be slow, but it should not be neglected in the QG&A roadmap. Learners find it easiest to express themselves when they can combine speech, gesture, and facial expressions, so these avenues should be explored.

Table 9 presents the tasks on the roadmap for interpreting learner answers. As brought up in the previous sections, it is important to develop content that adheres to the standards that evolve in ADL/SCORM and OLAMS.

Table 9: Interpreting Learner Answers

Tasks	Milestones			Measures
	3-year	5-year	10-year	
Developing and testing natural language understanding modules in MUC	<p>Develop additional standardized test sets and corpora more related to education/tutorial type dialog</p> <p>Development and Initial tests of NL methods on new standardized sets</p> <p>Continuing yearly MUC evaluation workshops</p>	NL understanding modules perform within 25% of human interpretation of answers	NL understanding modules perform within 10% of human interpretation of answers	3-8,11, MUC evaluation metrics
Developing and testing computational modules that interpret and evaluate lengthy verbal answers of learners	<p>Development and testing of providing feedback of individual components within lengthy verbal answers</p> <p>Modules able to segment and identify specific knowledge components within answers within 20% accuracy of human performance</p> <p>Modules able to provide overall scores within 10% of human accuracy across a range of domains</p>	<p>Modules able to segment and identify specific knowledge components within answers within 10% accuracy of human performance</p> <p>Modules able to provide overall scores at same accuracy as humans across a range of domains</p>	<p>Modules able to segment and identify specific knowledge components within answers at same accuracy of human performance</p> <p>Modules able to tie knowledge gaps to metatagged SCOs for feedback</p>	2, 3-8, 11
Evaluating the impact of learner answer interpretation facilities on learning	Understand the impact caused by the gap between perfect facilities and current state of the art, as well as comparison with control conditions	Studies lead to differentiation of results based on performance of individual understanding components	Systems designed to be robust to limitations of interpretation technology	1,2, 9, 10
Building software utilities and authoring tools for marking up documents in learning repository and NLP components	<p>Initial development of tools for tagging and segmenting content.</p> <p>Evaluation of automatic metatagging compared to human metatagging</p> <p>Performance within 30% of human performance</p>	<p>Continued development of tools for tagging and segmenting content</p> <p>Automatic Matching of content to pedagogical taxonomies and educational objectives</p> <p>Evaluation of automatic metatagging compared to human metatagging</p>	<p>Continued development of tools for tagging and segmenting content</p> <p>Automatic Matching of content to pedagogical taxonomies and educational objectives</p> <p>Evaluation of automatic metatagging compared to human metatagging</p>	1,11, conformity to ADL SCORM and OLAMS standards

		Performance within 10% of human performance and interface speeds up human performance	Performance at same level as that of human and interface speeds up human performance	
Maximizing the coverage of the landscape of questions (QSKP) in learning environments and scenarios	Initial repository of example questions covering landscape (both questions asked and answered) developed and linked to question taxonomy Examination and refinement of the distribution of questions across the QKPS and identification of most promising cells in the distribution for research	Collection of corpora of example questions across taxonomy for all 5 classes of learning environments is completed Question corpus annotated by hand for linguistic, knowledge, pedagogical information. (similar to efforts in table 8)	Complete repositories of example questions across taxonomy for additional new classes of learning environments	1
Developing and testing computational modules that interpret and evaluate input from visual and action modalities	Feasibility studies of interpretable modalities that most greatly impact learning and feedback Logging and integration of mouse actions	Initial development of modules to interpret/evaluate visual and action modalities Integration of pens, eyetrackers, gesture analysis, etc.	Comparison of methods of interpreting visual and action modalities. Methods contribute to improving overall system (e.g. verbal interpretation) by an additional 10%	3-8,11

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time

5.6 Collaborative Mixed-initiative Dialog

Some of the recent intelligent tutoring systems hold conversations with learners in natural language dialogs. Examples of these systems are AutoTutor (Graesser, Person Harter, & Tutoring Research Group, 2001), Why/AutoTutor and Why/Atlas (Graesser, VanLehn, Rose, Jordan, Harter, 2001), PACO (Rickel, Lesh, Rich, Sidner, & Gertner, 2002) and others (see Gratch, Rickel, Andre, Badler, Cassell, & Petajan, 2002). These systems present students' questions, interpret student contributions, and collaboratively construct answers in a turn-by-turn, mixed-initiative dialog. There is an attempt to comprehend whatever the student says by various modules in computational linguistics. AutoTutor is capable of evaluating the quality of learner contributions as well as graduate student research assistants, simply by using latent semantic analysis to compare student contributions (distributed over several turns) to an ideal expected answer. Some of these systems (AutoTutor, Why/AutoTutor, PACO) have animated conversational agents.

One important task in building a conversation system is to manage mixed initiative dialog. The computer tutor needs to be responsive to what the student says and also advance the conversation to meet pedagogical goals. The tutor for instance needs to produce the following dialog moves during each turn when responding to the student:

- 1) Ask the student major questions or present problems that will require major attention and conversation.
- 2) Respond to student assertions by giving positive, negative, or neutral feedback. This feedback can be in a variety of forms: verbal feedback without intonation, verbal feedback with intonation, facial expressions, a visual symbol on the interface, and so on.
- 3) Give hints or suggestions that help the student construct answers or solve problems.
- 4) Ask clarification questions when a student contribution is not understood by the computer system.
- 5) Ask diagnostic questions that diagnose whether the student understands a correct idea or has a misconception
- 6) Answer learner questions when asked.
- 7) Present a multimedia display that answers the learner's question.
- 8) Correct a learner's misconception.
- 9) Redirect a dead-end exchange by changing the topic or asking a new question.
- 10) Announce the limits of what the computer knows (e.g., *I don't understand*).
- 11) Shift gears when the learner changes the topic.
- 12) Signaling the learner when it is the learner's time to speak and when the tutor wants to keep the floor.

Mixed initiative tutorial dialog has been achieved, with surprising success, with finite state systems (Graesser, VanLehn, Rose, Jordan, & Harter, 2001), but these simple systems are far from perfect. The alternative is to dynamically plan the content of each

tutor plan by tracking and considering the dialog history, as well as the beliefs, desires, and intentions (BDI's) of the learner (Rickel et al., 2002). However, tutoring systems with dialog histories and BDI's are very difficult to develop and run the risk of being very slow.

One task is to evaluate the pedagogical quality and conversational smoothness of the questions, answers, and moves that the computer tutors present. This requires experts to evaluate the quality of the dialog by rating particular dialog moves on various dimensions of quality (Person, Graesser, Kreuz, Pomeroy & TRG, 2001). Alternative dialog moves are randomly selected so that expert judgments can be made on low-quality observations and compared with the computer moves; this permits the computation of recall, precision, F-measures, false alarms, and d' scores. An alternative method of evaluation is to conduct a bystander Turing test (Person, Graesser, & TRG, 2002). Dialog moves are generated by either a human or a computer tutor at random points in a dialog. A bystander decides whether each of these dialog moves is generated by a computer or a human. If the bystander cannot discriminate between a human and a computer tutor, then the computer is doing a good job simulating a human tutor. If not, the computer system needs to be improved.

As with humans, it is much more difficult to design computer systems that generate natural language than understand natural language. The natural language generation components in current tutorial dialog systems are currently at a very primitive state, but are improving (Moore & Wiemer-Hastings, 2003; Zukerman & Litman, 2002). Most systems have canned sentences, utterances, or paragraphs that get triggered by production rules. Others have template-driven semantic constructions that select a particular template and fill associated slots with referents that are available in the discourse history. A small number of systems dynamically plan sentences on the fly in a fashion that fits the constraints of what the system knows about the user's beliefs, desires, and intentions (Gratch et al., 2002; Rickel et al., 2002). However, these more sophisticated natural language generation systems are on very narrow applications and knowledge domains, such as train schedules, airline reservations, or spatial navigation. A dynamic generation of answers in natural language is one of the challenging problems that will be faced in the QG&A research agenda.

So far this section has addressed one-on-one tutoring between the computer and student. There is no reason to limit mixed initiative dialog to a two-part interaction between an expert and learner. It is possible to have agents that are peers, to have a group of agents (humans or avatars), to have a teacher interact with students, and to have other configurations of agents. Learners can model good question asking, question answering, and pedagogy by observing other agents (again, humans or avatars) that exhibit the sort of dialog moves and content that is desired. It is interesting to note that students rarely see inquisitive, skilled, learners who ask good questions because students rarely ask questions in the classroom. Students rarely observe good collaborative dialog because classroom environments are set up for teacher monologs more than dialog. It could be argued that most students rarely observe good pedagogy in the classroom. With these considerations in mind, substantial gains in inquiry and learning may occur when

students have the opportunity to view learning environments with multi-agent collaboration.

Table 10 presents the tasks that need to be completed when building systems with mixed initiative dialog. Some of the tasks and evaluation metrics will build on MUC and ARDA AQUAINT whereas others are unique for learning environments.

Table 10: Managing Mixed-initiative Dialog

Tasks	Milestones			Measures
	3-year	5-year	10-year	
Develop & evaluate dialog moves in mixed initiative dialog	Evaluation and comparison of existing systems. Systems give the <i>impression</i> of mixed initiative	Dialog move managers in real mixed initiative dialog, including allowing for interruptions etc.	Mixed initiative dialog using speech recognition and speech synthesis	1-8, 11, 12, bystander Turing test
Maximizing the coverage of the landscape of questions (QSKP) in learning environments and scenarios	Evaluation of taxonomies and scenarios (experimental psychology, corpus linguistics)	Taxonomies based on various scenarios and learning and evaluation	Broad coverage of question answering systems in different scenarios (incl. Multimedia)	1
Evaluating the impact of dialog facilities on learning	Currently available systems tested	Dialog facilities accommodate learning, incl. new technologies	Latest technologies implemented	9, 10
Building software utilities and authoring tools for marking up documents in learning content repository and NLP components	Systems use current tagging sets	Authoring tools address user modeling, domain, goals etc.	Implementation in overall framework, automated tagging process	1,11, conformity to ADL SCORM and OLAMS standards
Building software utilities and authoring tools for creating and modifying dialog management	Systems use current tagging sets	Authoring tools address user modeling, domain, goals, etc.	Implementation in overall framework, automated tagging process, mixed initiative dialog	1,11, conformity to ADL SCORM and OLAMS standards
Developing and testing natural language processing and dialog modules in MUC and ARDA AQUAINT	Implementation of existing systems in broader NLP framework, plus initial evaluation experiments	Framework language and dialog modules, e.g. include speech recognition and speech synthesis plus evaluation experiments	Latest ARDA AQUAINT, TREC and MUC systems, including audio and video, speech recognition and speech synthesis implemented, plus evaluation experiments	3-8,11, MUC and ARDA AQUAINT evaluation metrics
Developing and testing computational modules that scaffold simulation and visual media	Development, portability, testing effect of tools	Implementation into broader frameworks (NLP)	Systems direct learner to simulation and visual media when needed, interrupt when needed, etc.	3-8,11,12
Developing and testing learning environments with multiagent	Testing effect of (multi)agents	Development of multi agent collaboration	Implementation in overall framework	

The Learning Federation www.thelearningfederation.org

For more information contact: Kay Howell (khowell@fas.org)

collaboration				
---------------	--	--	--	--

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time; 12=learner confusion

5.7 User Modeling

A learning environment will be more effective if it is responsive to the abilities, knowledge, and other characteristics of the learner. There is a long history in psychology and education of measuring the learner's intelligence, achievement, aptitude, learning style, personality, and dozens of other traits. These assessments of general traits of the learner normally require the individual to take a psychometric test that ranges from 10 minutes to 5 hours, as in the case of the Armed Services Vocational Aptitude Battery (ASVAB) or the Scholastic Aptitude Test (SAT). If a profile of these test scores is available in the sample of learners, then these data could somehow affect the activities in the learning environment. If not, it is a bit cumbersome and distracting to put learners through a battery of tests before getting down to the business of helping them to learning the important material.

Researchers in psychology and education have also perfected the methods of designing tests to assess what a learner knows about a particular subject matter, however broadly or narrowly it is defined. Summative assessment occurs by administering tests on the material and computing 1 or more scores that reflect how well the learner has mastered the content. This normally occurs after the training is finished. In contrast, formative assessment occurs during the process of learning in a series of mini-evaluations that end up giving feedback to learners on what they know or don't know. This feedback presumably helps learners monitor their learning experience. The formative feedback can either be direct, as in the case of periodic on-line mini-tests, or be indirect, not appearing to be like a test at all. For example, modern intelligent tutoring systems (ITS) have indirect methods of inferring what each student knows by the nature of the student's decisions, answers, actions, and language. The ITS systems monitor their mastery of dozens or hundreds of skills, ideas, or misconceptions through the use of production rules (Anderson, Corbett, Koedinger, & Pellitier, 1995), Bayesian networks (VanLehn, 1990, 1996), latent semantic analyses (Graesser, Person et al., 2001), or some other diagnostic facility (Hunt & Minstrell, 1996). It is the indirect formative assessments that have attracted the interest of nearly everyone in the learning enterprise because these assessments are fine-grained, unobtrusive, and responsive to the learner. They allegedly provide the most learning gains. Indeed, the learning gains of ITSs have been extremely impressive (Corbett, 2001), approximately 1 sigma or letter grade, far better than alternative computer-based learning environments that have been evaluated. Some of the best ITSs available to date have been developed and evaluated at Carnegie Mellon and The University of Pittsburgh.

Broadly construed, user modeling is simply the process of collecting information about the user and having this information influence the activities of the learning environment. The user profile would include, but would not be limited to:

- a. Intelligence, technical knowledge, ability level of student

- b. Knowledge, experience, and background of student in the subject matter or collateral areas
- c. Success/failure on previous tests, evaluations, and assessments
- d. Success/failure of previous responses in learning session
- e. Success/failure of different learning approaches for this student
- f. Current evaluated ability of student
- g. Tailored learning tasks for this student
- h. Motivation and interest of the student
- i. Emotion of student at particular state
- j. Learning style (e.g., is student taking initiative)
- k. Learning disabilities
- l. Portfolio of written work and performance of student (student history)
- m. Mastery of ideas, beliefs, etc. associated with specific learning environments
- n. Existence of bugs, errors, and deep misconceptions associated with specific learning environments

There are a number of logistical and practical matters that need attention when a learner profile is monitored and used. First, there should be a normative distribution of values on these user variables, so a student can be compared with some reference set. Second, there should be informed consent from the student or parent that user modeling can proceed, so that there are no research ethics violations. Third, the data would presumably be passed on to a centralized database or to other instructors for future use.

One task for the QG&A road map is to decide what information should be included in the user profile and to build such data structures. In an effort to adhere to ADL/SCORM standards, there is a need to declare what metadata fields should be incorporated in XML. These fields would then be acknowledged by intelligent learning management systems so that the user profile is coordinated with the learning system. Additional tasks are shown in Table 11.

Table 11: Tasks: User Modeling

Milestones				
Tasks	3-years	5-years	10-years	Measures
Develop metadata fields in databases for user profiling and modeling	Metatagging schema for components and content	Prototypes of models embedded in some online learning environments. Results of studies that compare efficiency, utility, and validity of different learner models in different contexts	Prescriptions and standards for learner models exist and are widely adopted; Prescriptions for when to use more and less detailed learner models	11
Evaluating the reliability and validity of the variables in the user profile	Specifications for different types of scoring, analysis, and aggregation to generate mastery inferences	Demonstrations of complex configurations of multiple scoring of multiple task and response types and aggregation of data; Results of studies that establish the validity of the methods	Completely flexible scoring, analysis, and aggregation mechanisms provided by Web services	2,9,10
Integrate user profile with learning management systems	Specification of integration needs	Demonstration of integration with online learning systems, simulations, systems that store competency models; Analysis and reporting applications and services	Seamless interoperability of profiles and models and content with other systems and services	11, conformity to ADL SCORM and OLAMS standards
Evaluating the impact of user modeling on learning	A few hypotheses about the impact on learning, but need for more analysis	Some solid positive results; better understanding of how to adjust Q&A based on the user profile	Multiple solid results; evaluation implemented theories for different users, environments and scenarios	9, 10
Building software utilities and authoring tools for marking up databases	Development of particular utilities and tools toward integration (e.g. conformity)	Development of wide range of utilities and tools toward integration (e.g. conformity)	Development and implementation of variety of utilities and tools in QG&A	11, conformity to ADL SCORM and OLAMS standards

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time

5.8 Incorporating Learning Pedagogy

The design of a learning environment obviously depends on the underlying theory or philosophy of education and pedagogy. The pedagogical theory should have a direct impact on the information the system delivers, what information it collects from students, how it responds to students, and all levels of the mixed initiative dialog. For example, should the system provide the best factual answer or try to engage the student in a Socratic dialogue? Should the system refuse to answer immediately and let the student struggle a bit for a prescribed period? Or should the system immediately assist the student and correct misconceptions? Should the learner be part of a collaborative learning environment, or should the student learn independently?

The landscape of pedagogical theories is quite large. The most complex ideal theories include the Socratic method (Collins, 1985), modeling-scaffolding-fading (Collins, Brown, & Newman, 1989; Rogoff, 1990), reciprocal training (Palincsar & Brown, 1984), anchored situated learning (Bransford et al., 1991), error diagnosis and remediation (Sleeman & Brown, 1982; Hunt & Minstrell, 1996), frontier learning, building on prerequisites (Gagne, 1977), and sophisticated motivational techniques (Lepper, Woolverton, Mumme, & Gurtner, 1991). Graesser, Person and Magliano (1995) analyzed the presence of these pedagogical approaches in normal tutoring sessions with humans and found that the tutors do not use most of the ideal tutoring strategies. Tutors clearly need to be trained how to use sophisticated tutoring tactics because they do not routinely emerge in typical tutoring sessions. This analysis of the anatomy of human tutoring underscores the potential practical value of intelligent learning environments. If the computers can implement even a portion of the ideal tutoring strategies, then there should be substantial learning gains. Available meta-analyses of human tutoring (Cohen, Kulik, & Kulik, 1982) reveals learning gains of .42 standard deviation units whereas the estimate for ITS systems is 1.0 (Corbett, 2001) or higher.

Authoring tools need to be developed that support the particular pedagogical theories. The components of the authoring tool need to instantiate the theory and at the same time mesh with the other software models identified in this roadmap. Table 12 reviews the tasks that need to be completed.

Table 12: Learning Pedagogy

Tasks	Milestones			Measures
	3-year	5-year	10-year	
Evaluating the impact of pedagogical theories on learning	Overall evaluation of theories with additional evidence A few hypotheses about the impact of QG&A on learning, but need for more analysis	Recommendations for theories for different users, environments and scenarios Some solid positive results; better understanding of circumstances in which QG&A work	Multiple solid results; evaluation implemented theories for different users, environments and scenarios	9, 10
Building software utilities and authoring tools for implementing pedagogical theories	Utilities and tools based on selection of existing theories (development drives selection of theory)	Utilities and tools based on selection of best learning theories (theory drives development)	Conformity to standards, implementation in learning systems using a variety of interfaces	11, conformity to ADL SCORM and OLAMS standards
Maximizing the coverage of the landscape of questions (QSKP) in learning environments and scenarios	Few studies on pedagogical impact, limited to specific domains Coverage questions based on learning gains	A few multi-domain results, but still limited generality Coverage questions based on learning gains particular questions dependent on users, environments and scenarios	Several multiple-domain results Systems use history, user, environment, scenario system to select question that provides highest learning gains	1
Building software utilities and authoring tools that integrate other modules in QG&A	Development of particular utilities and tools toward integration (e.g. conformity)	Development of wide range of utilities and tools toward integration (e.g. conformity)	Development and implementation of variety of utilities and tools in QG&A	11, conformity to ADL SCORM and OLAMS standards

Measures: 1=incidence score; 2=interjudge agreement scores; 3=recall score; 4=precision scores; 5=f-measure; 6=hit rate; 7=false alarm (FA) rate; 8=d' score; 9=effect size; 10=variance of criterion measure explained by predictor measure; 11=response time

5.9 Linking People into the QG&A Systems

The design of most learning environments will involve people that play a variety of roles. So the learners will not normally be alone, but instead will be part of a society. There needs to be support for questions in learning communities, social construction of knowledge, situated cognition, distributed cognition, informal education, and lifelong learning. There is a large cast of individuals who contribute to the design and application of these learning environments. The society of individuals includes:

- a. The teacher or coach for the course
- b. Peers of the student
- c. The designer of the authorware
- d. Technical support that introduces the learner to the system and how to use it
- e. Support personnel on the help desk when the learning environment fails
- f. Subject matter experts who handle difficult questions either synchronously or asynchronously, by computer or by telephone
- g. Technical support that updates changes to the content and learning management system

The complete system will require a smooth integration, coordination, and timing of all of these individuals. Software will be designed, developed, and tested to facilitate these activities.

6 Launching and Stimulating the QG&A Research Agenda

Special steps may be taken to launch the QG&A research agenda. The first approach is proactive and top-down. A global architecture of the entire system will be designed and research groups will be assigned to work on particular components of the system. Each research team will need to precisely know what software will be designed, how it will be tested, what tasks will be completed, and what the milestones and deadlines are. One example of this approach is the DARPA funded Galaxy Communicator architecture, which established a standard for dozens of groups working on dialog management and speech recognition. Another example is the large scale research effort at the Institute for Creative Technologies, funded by the Army to act as a link between training systems in the Army and media developed in Hollywood. This top-down approach requires an early consensus of what to build, a strong project manager, and careful attention to deadlines and deliverables.

The second approach is proactive but more bottom up. There would be competitions on developing and testing systems in different categories of learning scenarios. This essentially is the Robot Wars and RoboCup model, where participants compete for the best system. Competitive approaches like these are known to dramatically increase interest in large numbers. The ground rules need to be established ahead of time on what the categories of systems are, what the minimal requirements are for entering a competition, and how performance will be measured. For instance, because of the diversity in tasks that have been defined in this Roadmap, systems may be evaluated on a subset of tasks. NIST could play a role in refereeing such a competition. This would be an extension of the TREC and MUC, except for the emphasis on learning environments and competition.

Whatever method is pursued in launching the QG&A agenda, it needs to be proactive and organized. We need to enable an interdisciplinary enterprise that brings together various groups, both within and across industry and academia – groups that have historically not worked together. These collaborations are necessary in order to take technology-enabled learning systems to the next level.

7 References

- Allen, J. (1995). *Natural language understanding*, Menlo Park, CA: Benjamin Cummings.
- Anderson, J.R., Corbett, A.T., Koedinger, K. & Pelletier, R. (1995). Cognitive tutors: lessons learning. *The Journal of the Learning Sciences*, 4, 167-207.
- Atkinson, R.K. (2002). Optimizing learning from examples using animated pedagogical agents. *Journal of Educational Psychology*, 94, 416-427.
- Baeza-Yates, R. & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Essex, England: Addison Wesley.
- Beck, I.L., McKeown, M.G., Hamilton, R.L., & Kucan, L. (1997). *Questioning the author: An approach for enhancing student engagement with text*. Delaware: International Reading Association.
- Belanich, J., Wisher, R. A., & Orvis (2003). *Web-based collaborative learning: An assessment of a question generation approach*. Technical Report 1133 of the United States Army Research Institute for the Behavioral and Social Sciences. Alexandria, VA
- Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive Domain*. New York: McKay.
- Bransford, J. D., Goldman, S. R., & Vye, N. J. (1991). Making a difference in people's ability to think: Reflections on a decade of work and some hopes for the future. In R. J. Sternberg & L. Okagaki (Eds.), *Influences on children* (pp. 147-180). Hillsdale, NJ: Erlbaum.
- Brown, A. L. (1988). Motivation to learn and understand: On taking charge of one's own learning. *Cognition and Instruction*, 5, 311-321.
- Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R. Harabagiu, S., Israel D. Jacquemin, C. Lin, C-Y, Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., Weishedel, R. (2001). *Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A)*. NIST.
- Card, S., Moran, T. & Newell, A. (1983). *The psychology of human-computer interaction*. Hillsdale, NJ: Erlbaum.

The Learning Federation www.thelearningfederation.org

For more information contact: Kay Howell (khowell@fas.org)

Carroll, J.M. (2000). Making Use: Scenarios and scenario-based design. In C. Paris, N. Ozkan, S. Howard & S. Lu (Eds.), *Interfacing reality in the new millennium. Proceedings of Oz CHI 2000*. (pp. 35-48). Canberra: Ergonomic Society of Australia.

Carroll, J.M. & Rosson, M.B. (1992). Getting around the task-artifact framework: How to make claims and design by scenario. *ACM Transactions on Information Systems*, 10(2), 181-212.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439-477.

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63, 1-49.

Ciardello, A. (1998). Did you ask a good question today? Alternative cognitive and metacognitive strategies. *Journal of Adolescent & Adult Literacy*, 42, 210-219.

Cohen, P.A., Kulik, J.A., & Kulik, C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal*, 19, 237-248.

Collins, A. (1985). Teaching reasoning skills. In S. Chipman, J. Segal, & R. Glaser (Eds.). *Thinking and Learning Skills. Vol. 2*. (pp. 579-586). Hillsdale, NJ: Erlbaum.

Collins, A. (1988). Different goals of inquiry teaching. *Questioning Exchange*, 2, 39-45.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Erlbaum.

Collins, A., Neville, P. & Bielaczyc, K. (2000). The role of different media in designing learning environments. *International Journal of Artificial Intelligence in Education*, 144-162.

Corbett, A.T. (2001). Cognitive computer tutors: Solving the two-sigma problem. *User Modeling: Proceedings of the Eighth International Conference, UM 2001*, 137-147.

Cote, N., Goldman, S.R., & Saul, E.U. (1998) Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, 25, 1-53.

Craig, S. D., Gholson, B., & Driscoll, D. (2002). Animated pedagogical agents in multimedia educational environments: Effects of agent properties, picture features, and redundancy. *Journal of Educational Psychology*. 94, 428-434.

Dillon, T.J. (1988). *Questioning and teaching: A manual of practice*. New York: Teachers College Press.

Dillon, J. T. (1984). Research on questioning and discussion. *Educational Leadership*, 42(3), 50-56.

The Learning Federation www.thelearningfederation.org
For more information contact: Kay Howell (khowell@fas.org)

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human Computer Interaction*. Upper Saddle River, NJ: Prentice Hall.

Dumais, S. T. (1993), "LSI meets TREC: A status report." In: D. Harman (Ed.), *The First Text REtrieval Conference (TREC1), National Institute of Standards and Technology Special Publication 500-207*, pp. 137-152.

Flammer, A. (1981). Towards a theory of question asking. *Psychological Research*, 43, 407-420.

Foltz, P.W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, 8, 111-128.

Gagné, Robert M. (1977). *The conditions of learning and theory of instruction*. New York: Holt, Rinehart & Winston.

Graesser, A.C., Hu, X., Person, N.K., Jackson, G. T., Toth, J. (2002). *Modules and information retrieval facilities of the human use regulatory affairs advisor (HURAA)*. The 7th annual world conference on E-learning in Corporate, Government, Healthcare, & Higher Education. AACE: Montreal, Canada.

Graesser, A. C., Langston, M. C., & Baggett, W. B. (1993). Exploring information about concepts by asking questions. In G. V. Nakamura, R. M. Taraban, & D. Medin (Eds.), *The psychology of learning and motivation: Vol. 29. Categorization by humans and machines* (pp. 411-436). Orlando, FL: Academic Press.

Graesser, A. C., & McMahan, C. L. (1993). Anomalous information triggers questions when adults solve problems and comprehend stories. *Journal of Educational Psychology*, 85, 136-151.

Graesser, A. C., & Olde, B. (in press). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*.

Graesser, J. Otero, & J. A. Leon (Eds.) (2002). *The psychology of science text comprehension*. Mahwah, NJ: Erlbaum.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, 31, 104-137.

Graesser, A. C., Person, N., Harter, D., & the Tutoring Research Group (2001). Teaching tactics and dialog in AutoTutor. *International Journal of Artificial Intelligence in Education*, 12, 257-279.

Graesser, A. C., Person, N., & Huber, J. (1992). Mechanisms that generate questions. In T. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems*. Hillsdale, NJ: Erlbaum.

Graesser, A. C., VanLehn, K., Rose, C., Jordan, P., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, 22, 39-51.

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the TRG (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8, 129-148.

Graesser, A.C., & Wisher, R.A. (2002). Question generation as a learning multiplier in distributed learning environments. *United States Army Research Institute for the Behavioral and Social Sciences, Technical Report 1121*.

Gratch, J., Rickel, J., Andre, J., Badler, N., Cassell, J., Petajan, E. (2002). Creating interactive virtual humans: Some assembly required. *IEEE Intelligent Systems*, 54-63.

Green, D. M. & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Harabagiu, S.M. Maiorano, S.J. & Pasca, M.A. (2002). Open-domain question answering techniques. *Natural Language Engineering*, 1, 1-38.

Hegarty, M., Narayanan, N. H. & Freitas, P. (2002). Understanding machines from multimedia and hypermedia presentations. In J. Otero, J. A. Leon & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 357-384). Mahwah, NJ: Erlbaum.

Hestenes, D., Wells, M. & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141-158.

HUB-4 Broadcast News Evaluation English Test Material (1999). Linguistics Data Consortium.

Hunt, E., & Minstrel, J. (1996). Effective instruction in science and mathematics: Psychological principles and social constraints. *Issues in Education: Contributions from Educational Psychology*, 2, 123-162.

Johnson, W. L., & Rickel, J. W., & Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. *International Journal of Artificial Intelligence in Education*, 11, 47-78.

Jurafsky, D., & Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, NJ: Prentice Hall.

Kass, A. (1992). Question-asking, artificial intelligence, and human creativity. In T. Lauer, E. Peacock, & A.C. Graesser (Eds.), *Questions and information systems* (pp. 303-360). Hillsdale, NJ: Erlbaum.

Kieras, D.E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, 8, 255-274.

King, A. (1989). Effects of self-questioning training on college students' comprehension of lectures. *Contemporary Educational Psychology*, 14, 366-381.

King A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, 31, 338-368.

Kintsch, E., Steinhart, D., Stahl, G. & LSA research group (2000). Developing Summarization Skills through the Use of LSA-Based Feedback. *Interactive learning environments*, 8, 87-109

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kintsch, W. (2001). Predication. *Cognitive Science*, 25, 173-202

Lamere, P., Kwok, P., Gouvea, E., Raj, B., Singh, R., Walker, W., Warmuth, M., Wolf, P. (2003). The CMU SPHINX-4 speech recognition system. *ICASSP*.

Landauer, T.K., Foltz, P.W., Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

Lehmann, F. (1992) (Eds.) *Semantic networks in artificial intelligence*. New York: Pergamon.

Lehnert, W. G. (1978). *The Process of Question-Answering*. Hillsdale, NJ: Erlbaum.

Lenat, D.B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38, 33-38.

Lepper, M.R., Woolverton, M., Mumme, D.L. & Gurtner, J-L. (1993). Motivational techniques of expert human tutors: lessons for the design of computer-based tutors. In S.P. Lajoie & S.J. Derry (Eds.), *Computers as cognitive tools*. Hillsdale: Erlbaum.

Linn, MC & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*. Mahwah, NJ: Erlbaum.

Litman, D. J., Walker, M. A., & Kearns, M. J. (1999). Automatic detection of poor speech recognition at the dialogue level. *Proceedings of the Thirty Seventh Annual Meeting of the Association of Computational Linguistics*, 309-316.

Louwerse, M.M., Graesser, A.C., Olney, A. & the Tutoring Research Group (2002). Good computational manners: Mixed-initiative dialog in conversational agents. In C. Miller, *Etiquette for Human-Computer Work. Papers from the 2002 Fall Symposium, Technical Report FS-02-02* (pp. 71-76).

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist*, 32, 1-19.

Mayer, R.E. (2002). *The promise of educational psychology (Vols. 1 & 2)*. Upper Saddle River, NJ: Prentice-Hall.

The Learning Federation www.thelearningfederation.org
For more information contact: Kay Howell (khowell@fas.org)

Miyake, N. & Norman, D.A. (1979). To ask a question one must know enough to know what is not known. *Journal of Verbal Learning and Verbal Behavior*, 18, 357-364.

Norman, D. A. & Draper, S. W. (Eds.) (1986). *User centered system design: New perspectives on human-computer interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Olde, B. A., Franceschetti, D.R., Karnavat, Graesser, A. C. & the Tutoring Research Group (2002). The right stuff: Do you need to sanitize your corpus when using Latent Semantic Analysis? *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 708-713). Mahwah, NJ: Erlbaum.

Otero, J., & Graesser, A.C. (2001). PREG: Elements of a model of question asking. *Cognition & Instruction*, 19, 143-175.

Palinscar, A. S., & Brown, A. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1, 117-175.

Pane, J.F., Corbett, A.T. and John, B.E. (1996). Assessing dynamics in computer-based instruction. *Proceedings of ACM CHI '96 Conference on Human Factors in Computing Systems*, 197-204.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York: Basic Books.

Pasca, M. A. & Harabagiu, S. M. (2001). High performance question/answering. *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2001)*, 366--374.

Pellom, B., Ward, W. & Pradhan, S. (2000). The CU Communicator: An architecture for dialogue systems", *International Conference on Spoken Language Processing (ICSLP)*, Beijing China.

Pressley, M., & Forest-Pressley, D. (1985). Questions and children's cognitive processing. In A.C. Graesser, & J.B. Black (Eds.), *The psychology of questions* (pp. 277-296). Hillsdale, NJ: Erlbaum.

Rickel, J., Lesh, N., Rich, C., Sidner, C. L. & Gertner, A. S. (2002). Collaborative discourse theory as a foundation for tutorial dialogue. *Intelligent Tutoring Systems*, 542-551.

Rogoff, B. (1990). *Apprenticeship in Thinking*. New York: Oxford University Press.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, 66, 181-221.

The Learning Federation www.thelearningfederation.org
For more information contact: Kay Howell (khowell@fas.org)

Rosson, M.B. & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction*. San Francisco: Morgan Kaufmann.

Scardamalia, M., & Bereiter, C. (1985). Fostering the development of self-regulation in children's knowledge processing. In S.F. Chipman, J.W. Segal, & R Glaser (Eds.). *Thinking and learning skills*, Vol. 2 (pp. 563-577). Hillsdale, NJ: Erlbaum.

Schank, R.C. (1986). *Explanation patterns: Understanding mechanically and creatively*. Hillsdale, NJ: Erlbaum.

Schank, R.C. (1999). *Dynamic memory revisited*. Cambridge: Cambridge University Press.

Schank, R.C., Ferguson, W., Birnbaum, L., Barger, J., & Greising, M. (1991). ASK TOM: An experimental interface for video case libraries. *Proceedings of the 13th Annual Conference for the Cognitive Science Society* (pp. 570-575). Hillsdale, NJ: Erlbaum.

Shneiderman, B. (1987). *Designing the user interface*. New York: Addison Wesley.

Schnotz, W., Bannert, M. & Seufert, T. (2002). Towards an integrative view of text and picture comprehension: Visualization effects on the construction of mental models.

Sleeman D. & J. S. Brown. 1982. Introduction: Intelligent Tutoring Systems. *Intelligent Tutoring Systems*, 1-10.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND Corporation.

Sparck Jones, K. & Willet, P (1997). *Readings in information retrieval*. San Francisco: Morgan Kaufmann.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions*. Cambridge, MA: MIT Press.

VanLehn, K. (1996). Conceptual and meta learning during coached problem solving. In C. Frasson, G. Gauthier, & A. Lesgold (Eds.), *ITS'96: Proceedings of the Third International Conference on Intelligent Tutoring Systems* New York: Springer-Verlag.

Voorhees, E. (2001). The TREC Question Answering Track. *Natural Language Engineering*, 7, 361-378.

Voorhees, E.M., Tice, D.M. (2000). Building a Question Answering Test Collection. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 200-207.

Webb, N.M., Troper, J.D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Experimental Psychology*, 87, 406-423.

The Learning Federation www.thelearningfederation.org

For more information contact: Kay Howell (khowell@fas.org)

Webber, B. (1988). Question answering. In S.C. Shapiro Ed.), *Encyclopedia of artificial intelligence*: Vol. 2 (pp. 814-822). New York: Wiley.

White, B., & Frederiksen, J. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction*, 16, 3-117.

Whittaker, S. (2003). Mediated communication. In A.C. Graesser, M.A. Gernsbacher, & S.A. Goldman, *Handbook of discourse processes*. Mahwah, NJ: Erlbaum.