Evaluating the Effectiveness of SCoT: A Spoken Conversational Tutor

Heather Pon-Barry, Brady Clark, Elizabeth Owen Bratt, Karl Schultz and Stanley Peters

Center for the Study of Language and Information
Stanford University
Stanford, CA 94305 USA
{ponbarry, bzack, ebratt, schultzk, peters}@csli.stanford.edu

Abstract. SCoT is a tutorial dialogue system that engages students in natural language discussions through a speech interface. The current instantiation, SCoT-DC, is applied to the domain of shipboard damage control—the task of containing the effects of crises (e.g. fires) that occur aboard Navy vessels. This paper describes a recent evaluation of SCoT-DC and presents preliminary results showing: (1) the effectiveness of SCoT-DC as a learning tool, and (2) that speech recognition technology is mature enough to support use in tutorial dialogue systems.

Keywords: Intelligent tutoring systems, spoken dialogue, evaluation, empirical results, speech technology, shipboard damage control

1 Introduction

There seems to be an assumption within the ITS community that the current state of speech technology is too primitive to provide effective tutorial interaction. Although some tutoring systems have been making use of spoken input for years (Aist & Mostow, 1997), only recently have researchers begun incorporating spoken input into dialogue-based tutoring systems (Clark et al., 2001; Litman & Silliman, 2004). The majority of tutorial dialogue systems currently rely on typed input from the student (e.g., Graesser et al., 2000; Evens et al., 2001; Heffernan & Koedinger, 2002).

SCoT, a Spoken Conversational Tutor, was developed in order to investigate the advantages of spoken language interaction in intelligent tutoring systems. While it has yet to be shown whether spoken tutorial dialogue systems can be more effective than typed tutorial dialogue systems, arguments involving access to prosodic and temporal information as well as multi-modality have been made in favor of spoken dialogue (Litman & Forbes, 2003; Pon-Barry et al., 2004). Furthermore, a recent study comparing spoken versus typed tutoring found significantly greater learning gains in the spoken condition when the tutor was a human, but little difference between the two conditions when the tutor was a computer (Litman et al., 2004)—suggesting that there is an advantage to spoken interaction, but that current tutorial dialogue systems are still far from human-like levels of sophistication.

In the winter and spring of 2004, we ran an experiment at Stanford University to evaluate the effectiveness of SCoT. In particular, the following two hypotheses were tested and confirmed:

(1) Tutorial interactions with SCoT-DC (the current instantiation of SCoT) will help students learn shipboard damage control
(2) Speech recognition, when combined with current technology for natural language understanding and dialogue management, is accurate enough to support effective spoken tutoring interactions

This paper is organized as follows. Section 2 gives an overview of how SCoT-DC works, Section 3 describes the experimental procedure, Section 4 presents empirical results, and Section 5 offers some conclusions.

2 System Overview

SCoT-DC, the current instantiation of the SCoT tutoring system, is applied to the domain of shipboard damage control. Shipboard damage control refers to the task of containing the effects of fires, floods, explosions, and other critical events that can occur aboard Navy vessels. Students carry out a reflective discussion with SCoT-DC after completing a problem-solving session with DC-Train (Bulitko & Wilkins, 1999), a fast paced, real time, speech-enabled training environment for damage control. Figure 1 shows a screenshot of DC-Train.

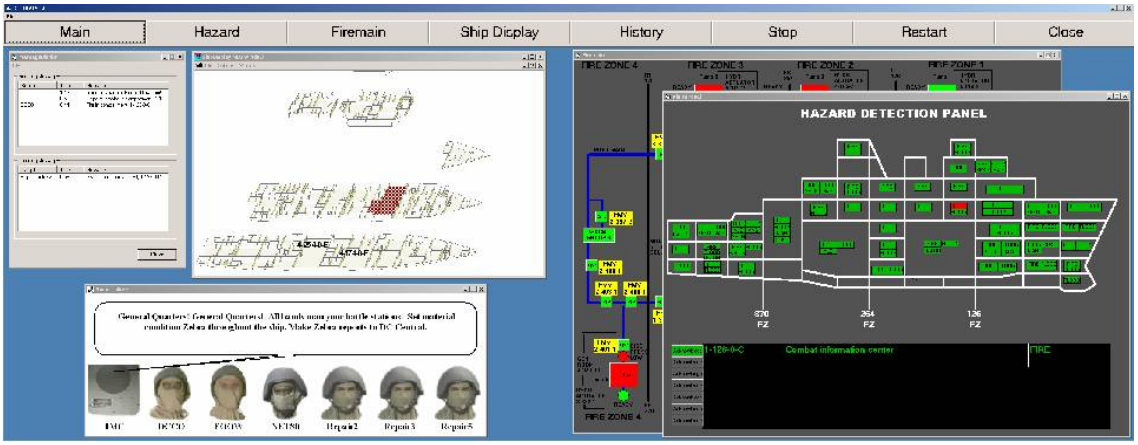

Figure 1. DC-Train Simulator

Figure 2 shows a screenshot of SCoT-DC. The window on the right depicts the multiple decks of the ship; the window in the bottom left corner contains a history of the tutorial dialogue as well as buttons for starting the tutor; and the window in the upper left corner is the *common workspace*—a space where both the student and the tutor can zoom in or out, and select (i.e. point to) compartments, regions, or bulkheads (lateral walls in the ship).

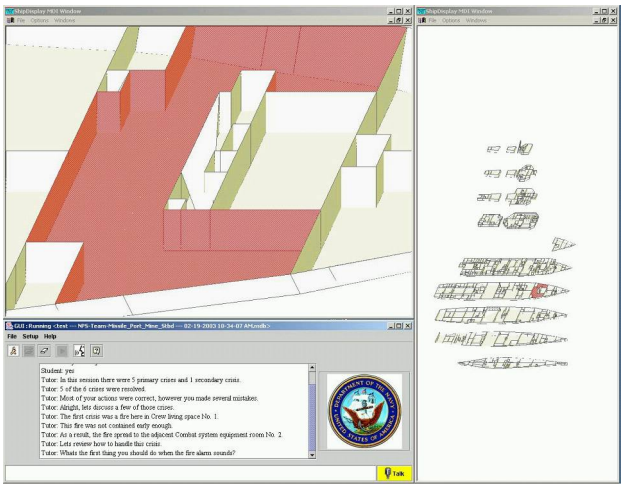

Figure 2. SCoT-DC Tutor

SCoT is developed within the Architecture for Conversational Intelligence (Lemon et al., 2002), a general purpose architecture which supports multi-modal, mixed-initiative dialogue. The version of SCoT used in these experiments consists of three separate components: a dialogue manager, a tutor, and a set of natural language tools. These three components are described briefly in sections 2.1 through 2.3. A more detailed description can be found in Schultz et al. (2003).[1] Figure 3 is the overall architecture of our system.

[1] Also, a downloadable video is available at www-csli.stanford.edu/semlab/muri/November2002Demo.html

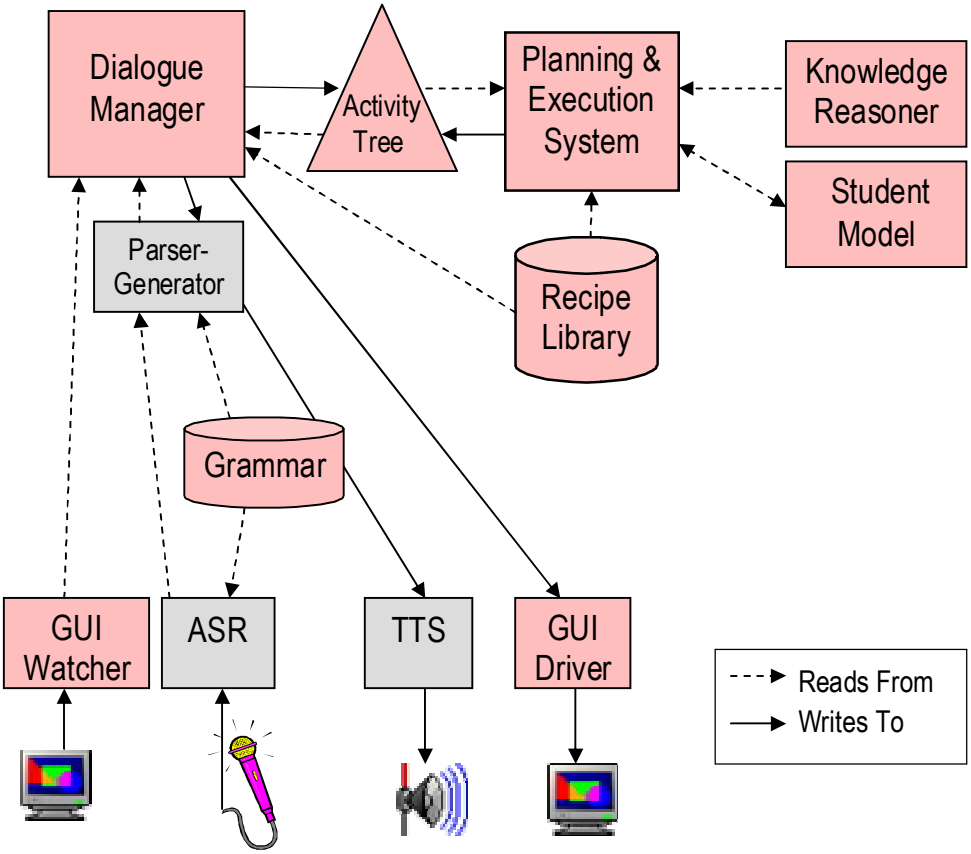

Figure 3 SCoT-DC Tutor architecture.

2.1 Dialogue Manager

The dialogue manager mediates communication between the system and the user by handling aspects of conversational intelligence such as turn management and coordination of multi-modal input and output. It contains multiple dynamically updated components—the two main ones are (1) the *dialogue move tree*, a structured history of dialogue moves, and (2) the *activity tree*, a hierarchical representation of the past, current, and planned activities initiated by either the tutor or the student. In SCoT, each activity initiated by the tutor corresponds to a tutorial goal; the decompositions of these goals are specified by activity recipes contained in the *recipe library* (see section 2.2).

2.2 Tutor

The tutor component contains the tutorial knowledge necessary to plan and carry out a flexible and coherent tutorial dialogue. It is implemented as a set of domain-general tutorial activities (the recipe library) and a system for planning and executing these activities. The *recipe library* contains activity recipes that specify how to decompose a tutorial activity into other activities and low-level actions. The *planning and execution system* uses information from external knowledge sources to (1) generate an initial plan at the start of the dialogue, (2) decide when and how to revise these plans during the dialogue, (3) classify student utterances, and (4) decide how to respond to student answers, i.e., which recipe to put on the activity tree. A high-level activity recipe for discussing a partially-correct answer is shown in Figure 4, followed by the resulting dialogue in Figure 5 (dialogue turns resulting from recipe are shown in boldface). The recipe in Figure 4 decomposes the activity of discussing a partially-correct answer into four sub-activities (acknowledge, hint, appraise, and generalize). The tutor places these activities on the activity tree, and the dialogue manager begins to execute their respective recipes.

Goal
- Discuss partially-correct answer/error of omission

Preconditions
- Student Answer (from dialogue) = Partially Correct
- Student Action (from DC-Train) = Error of Omission

Recipe Body
- Acknowledge: *Muted Praise*
- Hint: *Partially Correct Hinting Sequence*
- Appraise: *State Error of Omission*
- Generalize: *Comment on Repeated Errors*

Figure 4. Activity recipe for discussing a problem solving sequence

Tutor: All right, suppose the fire has been overhauled.
Tutor: Let's now discuss the resulting smoke. [displays grey color in compartment].
Tutor: What should you do first?
Student: Send repair two to desmoke the compartment.
Tutor: Almost.
Tutor: You should send a different repair team.
Tutor: Which team should you send instead?
Student: Repair five.
Tutor: Yes, exactly. You should order repair five to desmoke the compartment.
Tutor: Repair 2 was overtasked, so repair 5 should have been sent.
Tutor: But you sent the wrong repair team during the session.
Tutor: This is a mistake you made more than once.

Figure 5. Tutorial dialogue resulting from activity recipe in Figure 4

2.3 Natural Language Components

The natural language components which make the spoken dialogue possible include a bi-directional unification grammar and off-the-shelf tools for automatic speech recognition and text-to-speech synthesis. Incoming student utterances are handled by SCoT in the following way. First, the utterance is recognized using Nuance[2] speech recognition, which uses a grammar compiled from a Gemini natural language understanding grammar. Gemini (Dowding et al., 1993) translates word strings from Nuance into logical forms, which the dialogue manager interprets in context and routes to the tutor. The system responds to the student via a FestVox[3] limited domain synthesized voice.

3 Experiment

This experiment was designed to test the hypothesis that SCoT-DC will help students learn damage control. In order to test this hypothesis, we divided the tutorial content into three knowledge areas: sequencing, boundaries, and jurisdiction. Sequencing refers to giving orders for actions in response to crises (e.g. fires, floods) at the correct times. Setting boundaries refers to the task of correctly specifying six parameters that determine the location of the bulkheads (upright partitions that separate ship compartments) that need to be cooled or sealed to prevent a crisis from spreading. Jurisdiction refers to the task of giving orders to the appropriate personnel on the ship—personnel are assigned to different regions such as forward, aft, and midship.

[2] http://www.nuance.com
[3] http://festvox.org

3.1 Participants

Thirty native English speakers were recruited to participate in this experiment (16 male, 14 female). All subjects were novices in the domain of damage control, twenty-nine had no prior experience in dialogue system studies.

3.2 Experimental design

Subjects were randomly assigned to three groups. All groups ran through the same four DC-Train scenarios (which increased in difficulty). Between each DC-Train session, all groups received tutoring in one of the three knowledge areas (sequencing, boundaries, and jurisdiction), but at different times. For example, group I received tutoring on sequencing between scenario 1 and scenario 2, group II received tutoring on sequencing between scenario 3 and scenario 4, and group III received tutoring on sequencing between scenario 2 and scenario 3. This allowed us to separate learning gains due to the tutorial interaction from learning gains due to practice alone. In this way, each subject served as their own control, and all groups served as a comparison for each other. Table 1 shows the layout for each group.

Subject Group	**DC-Train Session**	**SCoT-DC Tutoring**	**DC-Train Session**	**SCoT-DC Tutoring**	**DC-Train Session**	**SCoT-DC Tutoring**	**DC-Train Session**
I	Scenario 1	*Sequencing*	Scenario 2	*Boundaries*	Scenario 3	*Jurisdiction*	Scenario 4
II	Scenario 1	*Boundaries*	Scenario 2	*Jurisdiction*	Scenario 3	*Sequencing*	Scenario 4
III	Scenario 1	*Jurisdiction*	Scenario 2	*Sequencing*	Scenario 3	*Boundaries*	Scenario 4

Table 1. Experiment design

Learning was measured in two ways. Firstly, general knowledge was tested in the form of a multiple-choice pre-test and a post-test. Secondly, and crucially, quantitative performance measures were drawn from each of the four DC-Train scenarios. Based on the logfiles of each scenario, students were scored for their performance in sequencing, boundaries, and jurisdiction.

3.3 Procedure

The experimental procedure is illustrated below in Table 2. Steps 4 through 10 (shown in boldface) constitute the main body of the experiment and correspond to the steps listed in Table 1. In addition to these main steps, all subjects went through an interactive multimedia introduction to (1) familiarize them with DC-Train and basic damage control knowledge, and (2) give them practice using the speech recognition interface. After the multimedia introduction, subjects took a 20 question multiple-choice pre-test, and had one practice DC-Train session. Following the main body of the experiment, subjects took a 20 question post-test and filled out a questionnaire. The total duration of the experiment was roughly three hours per subject.

Step 1	Multimedia Introduction	30-40 min
Step 2	Pre-test	5-10 min
Step 3	Practice DC-Train session	10 min
Step 4	**DC-Train session 1**	15 min
Step 5	**Tutoring**	< 15 min
Step 6	**DC-Train session 2**	15 min
Step 7	**Tutoring**	< 15 min
Step 8	**DC-Train session 3**	15 min
Step 9	**Tutoring**	< 15 min
Step 10	**DC-Train session 4**	15 min
Step 11	Post-test	5-10 min
Step 12	Questionnaire	< 5 min

Table 2. Experiment Procedure

4 Results

4.1 Hypothesis (1) – Effectiveness of Tutoring

Students showed improvement both on the written test and on their performance in the DC-Train scenarios. Every student earned a higher score on the post-test than on the pre-test, and the mean post-test score (84%) was significantly higher (Paired-Samples T Test: $p = 0.000$) than the mean pre-test score (67%). Looking at performance in the DC-Train simulator, students were performing better in their fourth session with the simulator than in the first for two of the three knowledge areas (sequencing and boundaries). These performance gains are shown in Figure 6.

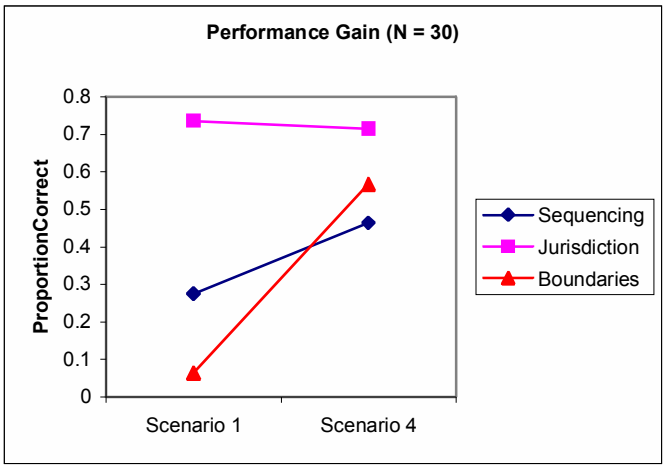

Figure 6. DC-Train performance gains

The lack of improvement in the area of jurisdiction may be due to a ceiling effect—because initial performance in jurisdiction was already very high. However, there are clear performance gains in the areas of sequencing and boundaries. This leads to the question of whether these gains should be attributed to the SCoT-DC tutoring or just to improvement over time with practice on the DC-Train simulator.

Because all three groups received tutoring in all three knowledge areas (at different times), we can separate gains due to SCoT tutoring from gains due to practice alone. We do this by comparing, for a particular knowledge area, performance gains across sessions with tutoring on that knowledge area to performance gains across sessions with tutoring on some other knowledge area. For example, consider the graphs in Figure 7. In Figure 7a, the left column depicts the average gain across two DC-Train scenarios between which the student received tutoring on sequencing (i.e. for subject group I: between scenario 1 and scenario 2, for subject group II: between scenario 3 and scenario 4, for subject group III: between scenario 2 and scenario 3). The right column depicts the average gain across sessions between which the student received tutoring on either boundaries or on jurisdiction. Figures 7b and 7c show the same data, but for boundaries and jurisdiction respectively.

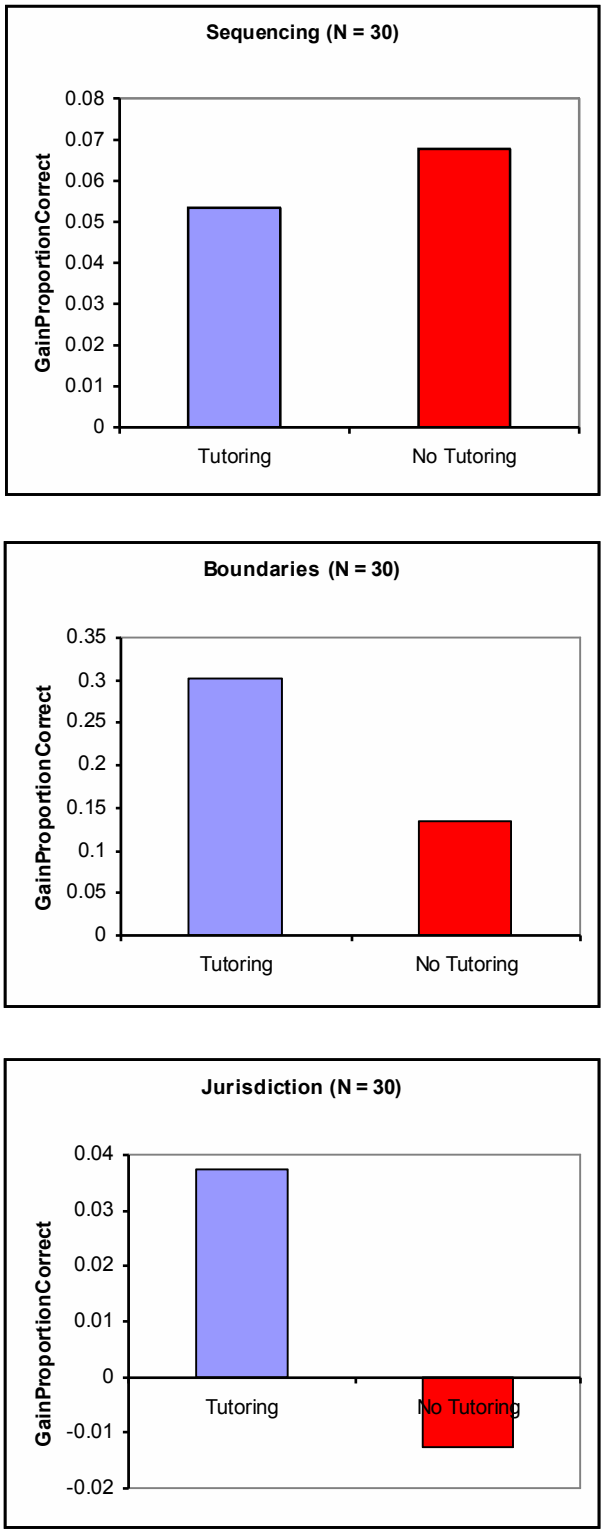

Figure 7. Gains due to tutoring vs. gains due to practice for (a) sequencing, (b) boundaries, and (c) jurisdiction

Figure 8 shows the average gains due to tutoring and due to practice (no tutoring) across all three knowledge areas. On average, gains in performance after being tutored in a particular knowledge area are over twice as high as gains in performance after being tutored in some other knowledge area.

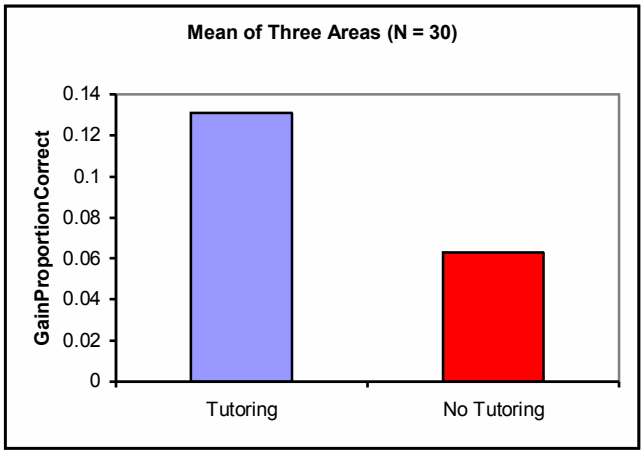

Figure 8. Gains due to tutoring vs. gains due to practice for all three knowledge areas

We have already seen that jurisdiction performance did not improve much with use of the system, perhaps because it started at such a high level, so the small difference seen in Figure 7c between the means of 0.04 (stdev=0.40) and -0.01 (stdev=0.21) in jurisdiction performance may not be as important to consider as the areas in which students did improve.

In the area of boundaries, the mean performance gain with tutoring (0.30) is close to one standard deviation above the performance gain with tutoring (0.13, stdev=0.21). This statistic gives the clearest evidence of the benefit of tutoring combined with simulator practice.

The fact that in the area of sequencing subjects had smaller average performance gains with tutoring (0.05, stdev=0.27) than those without tutoring (0.06, stdev=0.10) seems to argue against the effectiveness of the tutor. However, on examining the performance of each subject group separately, an interesting pattern emerges. Subjects who received sequencing tutoring first (group I) showed larger gains in sequencing in the following DC-Train scenario than they did in their subsequent scenarios (0.13 with tutoring, 0.03 with no tutoring). However, the other two subject groups, who received sequencing tutoring later, did not appear to benefit from the tutoring (group II: -0.01 with tutoring, 0.12 with no tutoring; group III: 0.04 with tutoring, 0.05 with no tutoring). A possibile explanation is that the subject matter of sequencing is fundamental to performance on DC-Train and that it is critical to be tutored on it early on. Allowing students to practice their mistakes before reviewing the correct actions may lead them into habits that are hard to unlearn. We hope that future experiments may clarify whether this explanation holds.

4.2 Hypothesis (2) – Effectiveness of Speech Interface

We collected the following statistics on speech recognition performance and on speech quantity:

- percentage of words correctly recognized
- percentage of sentences recognized with no word errors
- percentage of sentences rejected by the speech recognizer
- mean length of utterance in words
- mean number of utterances per session

Twenty speakers have been analyzed so far, though for several statistics only sixteen cases ended up meeting all the conditions necessary for the analysis.

Our main hypothesis was that speech recognition is accurate enough to support effective spoken tutoring interactions. We predicted that speech recognition performance would correlate with improvements on the written tests and with performance in simulator. We found no correlation with either one. For percentage words

correct, the correlation with test score gains is -.038, with a significance of .890 (N=16), and the correlations with performance in sequencing, boundaries, and jurisdiction were -.303, .216, and .016 respectively. Figure 9 shows a scatter-plot of average percent words correct versus the gains from the pre-test to the post-test.

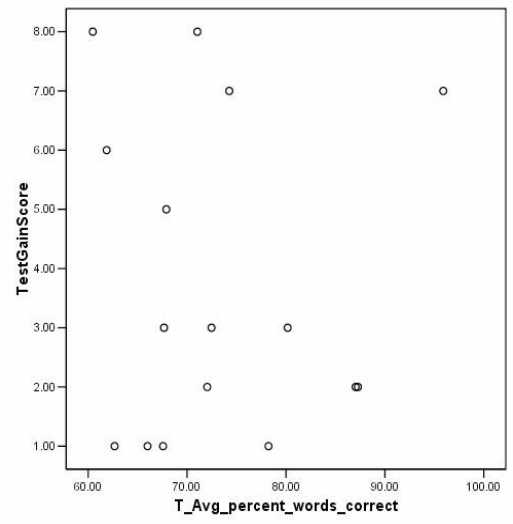

Figure 9. Scatter-plot of speech recognition success vs. gains in test scores.

Although our prediction was not validated, this result is highly relevant because it suggests that poor speech recognition does not diminish learning. Students who had only 60% of their words recognized correctly showed learning gains comparable to students who had 95% of their words recognized correctly. Even though speech recognition is far from perfect, students being tutored by SCoT-DC improved regardless of the number of misrecognitions they encountered. This result is in line with recent findings reported in (Litman et al., 2004).

While speech recognition accuracy did not affect learning, it did affect the student's desire to use the system regularly. As part of the questionnaire, subjects were asked to rate the following statement on a 1-7 Likert scale (1 = strongly disagree, 7 = strongly agree):

> "Based on my experience using this tutoring system, I would like to use this kind of automated tutoring system regularly."

Subjects said they would like to use this kind of system again when they had a higher percentage of sentences with no errors (Pearson correlation = .557, significant at the 0.05 level (2-tailed)). However, there was no correlation with percentage words correct, and paradoxically, high rejection rates also correlated with more desire to use the system (Pearson correlation = .629, significant at the 0.01 level (2-tailed)).

4.3 Future Analysis

In the future, we are interested in adding semantic error rate to the speech performance statistics, to see how many of the speech recognition errors made a difference in the interpretation the system assigned to the sentence. Semantic errors are the subset of word and sentence errors that result in the system misunderstanding the user (assigning an incorrect logical form to the utterance). For example, if the student says "send repair five to set fire boundaries" and the recognition hypothesis is "send repair five to set fire boundary", it would be a word/sentence error but not a semantic error. However, if the recognition hypothesis was "send repair five to check the fire" it would be both a word/sentence error and a semantic error. The interplay between word error and semantic error is discussed in Wang et al. (2003).

We are also interested in examining speech characteristics that may reflect a student's level of certainty in their answer such as: speech rate, pauses, filled pauses ("um" and "uh"), disfluencies ("se- set boundaries"), hedges ("I guess", "probably", "maybe"), and latency (the delay between the end of the system's question and the start

of the student's response). We would like to see if any of these metrics can be used in updating the student model and selecting appropriate tutorial tactics.

5 Conclusions

These results demonstrate that SCoT-DC's tutoring on the three knowledge areas was effective. Subjects who started off knowing nothing about the domain learned a surprising amount about shipboard damage control. In just three hours, their performance with the damage control simulator (voice-enabled DC-Train) increased to a level of 51% fully correct actions and their scores on a written test about damage control rose to 84% correct.

Additionally, the initial results demonstrate that speech recognition technology is mature enough to support usable instructional technology. And, the remaining room for improvement in speech technology does not overwhelm the pedagogical virtues and shortcomings of instructional simulators and intelligent tutors.

Acknowledgements

This work is supported by the Office of Naval Research under research grant N000140010660, a multidisciplinary university research initiative on natural language interaction with intelligent tutoring systems. Further information is available at http://www-csli.stanford.edu/semlab/muri.

References

1. Aist, G., & Mostow, J. (1997). A time to be silent and a time to speak: Time-sensitive communicative actions in a reading tutor that listens. *AAAI Fall Symposium on Communicative Actions in Humans and Machines*, Boston, MA.
2. Anderson, J. R. (1993). *Rules of the mind.* Hillsdale, NJ: Erlbaum.
3. Bulitko, V., & Wilkins., D. C. (1999). Automated instructor assistant for ship damage control. In *Proceedings of AAAI-99*.
4. Clark, B., Fry, J., Ginzton, M., Peters, S., Pon-Barry, H., & Thomsen-Gray, Z. (2001). A Multimodal Intelligent Tutoring System for Shipboard Damage Control. In *Proceedings of 2001 International Workshop on Information Presentation and Multimodal Dialogue (IPNMD-2001).* Verona, Italy. 121-125.
5. Dowding, J., Gawron, M., Appelt, D., Cherny, L., Moore, R., and Moran, D. (1993). Gemini: A natural language system for spoken language understanding. In *Proceedings of ACL 31*.
6. Evens, M., Brandle, S., Chang, R., Freedman, R., *et al*. (2001). CIRCSIM-Tutor: An Intelligent Tutoring System Using Natural Language Dialogue. In *Proceedings of the Twelfth Midwest AI and Cognitive Science Conference, MAICS 2001*.
7. Graesser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & the Tutoring Research Group. (2000). AutoTutor: a simulation of a human tutor. *Journal of Cognitive Systems Research, 1,* 35-51.
8. Heffernan, N., & Koedinger, K. (2002). An Intelligent Tutoring System Incorporating a Model of an Experienced Human Tutor. In *Proceedings of the 6th International Conference, ITS 2002*.
9. Lemon, O., Gruenstein, A., & Peters, S. (2002). Collaborative activities and multitasking in dialogue systems. In C. Gardent (Ed.), *Traitement Automatique des Langues (TAL, special issue on dialogue), 43(2),* 131-154.
10. Litman, D., & Forbes, K. (2003). Recognizing Emotions from Student Speech in Tutoring Dialogues. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, St. Thomas, Virgin Islands.
11. Litman, D., & Silliman, S. (2004). ITSPOKE: An Intelligent Tutoring Spoken Dialogue System. In *Proceedings of the Human Language Technology Conference: 4th Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL)* (Companion Proceedings), Boston, MA.
12. Litman, D., Rosé, C., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2004). Spoken Versus Typed Human and Computer Dialogue Tutoring. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems (ITS)*, Maceió, Brazil.
13. Pon-Barry, H., Clark, B., Schultz, K., Bratt, E., & Peters, S. (2004). Advantages of Spoken Language Interaction in Tutorial Dialogue Systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems*, Maceió, Brazil.
14. Schultz, K., Bratt, E., Clark, B., Peters, S., Pon-Barry, H., & Treeratpituk, P. (2003). A Scalable, Reusable Spoken Conversational Tutor: SCoT. In *Proceedings of the AIED 2003Workshop on Tutorial Dialogue Systems: With a View Towards the Classroom*.
15. Wang, Y., Acero, A., & Chelba, C. (2003). Is Word Error Rate a Good Indicator for Spoken Language Understanding Accuracy. In *Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding.* St. Thomas, Virgin Islands.