WIZARDS' DIALOGUE STRATEGIES TO HANDLE NOISY SPEECH RECOGNITION

Tiziana Ligorio[1], Susan L. Epstein[1 2], Rebecca J. Passonneau[3]

[1]Department of Computer Science, The Graduate Center of The City University of New York
[2]Department of Computer Science, Hunter College of The City University of New York
[3]Center for Computational Learning Systems, Columbia University
tligorio@gc.cuny.edu, susan.epstein@hunter.cuny.edu, becky@cs.columbia.edu

ABSTRACT

This paper reports on a novel approach to the design and implementation of a spoken dialogue system. A human subject, or wizard, is presented with input of the sort intended for the dialogue system, and selects from among a set of pre-defined actions. The wizard has access to hypotheses generated by noisy automated speech recognition and queries a database with them using partial matching. During the ambitious study reported here, different wizards exhibited different behaviors, elicited different degrees of caller affinity for the system, and achieved different degrees of accuracy on retrieval of the requested items. Our data illustrates that wizards did not trust automated speech recognition hypotheses when they could not lead to a correct database match, and instead asked informed questions. The wealth of data and the richness of the interactions are a valuable resource with which to model expert wizard behavior.

Index Terms— spoken dialogue systems, Wizard of Oz study, corpus resources

1. INTRODUCTION

In the design of a spoken dialogue system (*SDS*), a Wizard-of-Oz study offers a window into human expert behavior and supports learning a model of expertise. In such a study, a human subject (the *wizard*) is presented with real or simulated automated speech recognition (*ASR*) output, and her actions in response are recorded [12, 15, 20]. *Voice search* allows a wizard to query a backend directly with ASR output, and receive returns ranked by a similarity score [14]. The thesis of this work is that a study with an embedded wizard who uses voice search will produce a rich and novel corpus that exhibits varied performance among wizards and callers. This paper describes the collection of such a corpus of caller-wizard interactions.

In other work, wizards who had difficulty interpreting ASR (*non-understanding*) tried to continue their task in ways other than clarifying or repeating the utterance [15, 20]. The corpus described here highlights the alternatives wizards used when they were uncertain about what the caller had said. Our wizards worked to interpret the caller's request, given noisy ASR, voice search, and a large set of pre-specified questions derived from prior work. The principal result of this study is that two very different wizard strategies achieved similar success. In one approach, wizards were confident in their own assessment of the hypotheses' accuracy and the relevance of database returns. In the other, wizards asked more questions, sought confirmation more often, and had lengthier dialogues that were not necessarily more accurate but gave callers a sense of greater understanding and progress. The least successful wizard strategies differ from both approaches. This data will initially be used to train models of successful behavior to improve the SDS. The corpus, to be released in 2011, can support many other investigations.

The next section discusses background and motivation for this experiment. Subsequent sections describe our domain of investigation and experimental design, and provide a preliminary analysis of the collected corpus. The final section discusses how we will apply this important resource.

2. MOTIVATION AND BACKGROUND

This work seeks to elicit strategies that will serve well with the range of ASR performance common in fielded dialogue systems, a word error rate (*WER)* at best near 30%-35% and as high as 70% [11]. An effective SDS should minimize both misunderstandings and non-understandings. One way to address this goal is to aim for high accuracy in database retrieval despite high WER. The need to correct the system's misunderstandings, however, can frustrate the caller, and such attempts are more poorly recognized than non-correction utterances [6]. For non-understanding, re-prompting the caller for the same information often fails when hyperarticulation results in similar, or even worse recognition. Rather than frustrate the caller, wizards often use more creative ways to re-elicit the same information — they use contextual information and confirm that some communication has occurred.

In related work, wizards given ASR output performed surprisingly well despite a high WER [15]. Although dialogues about finding directions had a WER of 42%, misunderstanding occurred only 5% of the time, and partial understanding and non-understanding 20% of the time each. Rather than

signal non-understanding, wizards continued a route description, asked a task-related question or requested a clarification. Despite the high percentage of partial and non-understandings, users reported that they were well understood by the system. A dialogue study for a multimodal MP3 player application simulated noisy transcription by word deletion, and varied task difficulty by deletions of between 20% and 50% [12]. It also introduced lexical ambiguities in the database to elicit different kinds of clarification strategies. In the noisy condition, wizards asked for clarifications about twice as often as occurred in similar human-human dialogue. Another study of dialogues for tourist requests also artificially varied WER [20]. It reported that, under medium WER, task-related questions led more often to full understanding than did an explicit signal of non-understanding.

In an earlier study, we provided context for ASR disambiguation through voice search [7, 8]. Subjects queried a database of book titles and then selected the correct title from among as many as 10 returns with the highest match scores. (Matching is further described in Section 4.) In 4172 title cycles with high (71%) WER, voice search returned a list of more than one title to choose from 53.26% of the time, and otherwise returned a single, high-scoring candidate. When a title appeared among the search results, the subject either identified it with confidence (26.53%), identified it with some uncertainty (68.72%), or gave up (4.75%).

During a full dialogue the subject might have requested clarification on the uncertain identifications. Although voice search can improve recognition [17], there will always be a residue of cases where the input is so noisy that voice search fails. In those cases, models of how wizards disambiguate among voice search returns or use them to ask informed questions can be used to further improve the system. We were able to predict wizard behavior with accuracy as high as 82.2% from decision trees learned on a combination of system and session features recorded during the experiment. Linear and logistic regression models achieved comparable accuracy. These results motivated the experiment reported here, where, in full dialogues, wizards could use voice search and ask questions to disambiguate noisy ASR.

3. DOMAIN OF INVESTIGATION

The Wizard-of-Oz study reported here models book order transactions at the Andrew Heiskell Braille and Talking Book Library, a branch of the New York Public Library and part of the National Library Service (*NLS*). Patrons receive a monthly catalogue of new and popular library holdings, with book titles, authors, and catalogue numbers. Patrons' requests are handled by telephone, and received by mail. Given increasing caller volume and limited staff, Heiskell and other NLS libraries could benefit greatly from an SDS that automates some borrowing requests.

The baseline SDS *CheckItOut* was implemented within the Olympus/Ravenclaw dialogue system architecture [4]. Olympus has thus far supported about a dozen substantial dialogue systems in different domains, including Let's Go Public! [11]. Among the Olympus components, we chose PocketSphinx for speech recognition, and used freely available acoustic models of Wall Street Journal dictation speech, adapted with about 8 hours of spontaneous speech for our domain. The speech data for the current experiment has not yet been transcribed, but a sample of 315 transcribed utterances with the same recognition settings and 6 speakers suggest that the WER was about 50%.

For natural language understanding, we used Phoenix, a robust, semantic parser [19]. Phoenix produces one or more semantic frames per input ASR string. When some words cannot be parsed, a frame may be a discontinuous sequence of slots. Each slot has an associated context-free grammar (*CFG*), and corresponds to a concept. To manage the large vocabulary and rich syntax of book titles, we parsed the entire 71,166-title database with a large-coverage dependency grammar [1], and then mapped the parses to the CFG format Phoenix requires. The remaining Phoenix productions were generated by hand. The grammar and language models for book titles were built from 3000 randomly-selected book titles. We also used the Apollo interaction manager [10] to detect utterance boundaries using information from speech recognition, semantic parsing and utterance-level confidence, as measured by the Helios confidence annotator [2].

CheckItOut's backend accesses a sanitized version of Heiskell's database of 5028 active patrons, plus its full book and transaction databases for 71,166 titles and 28,031 authors. Although titles and author names include 54,448 distinct words, CheckItOut's vocabulary, as reflected by its grammar and language model, consists of only 8,433 words. For the experiment described here, a wizard server replaced the dialogue manager. Runtime data from many components supported the construction of models of wizard behavior that can be used to improve the baseline system.

4. EXPERIMENTAL DESIGN

Ten callers (5 male, 5 female) each made 15 calls to each of 6 wizards (3 male, 3 female), for a total of 900 calls. Wizards and callers were recruited by email and flyers to students at Hunter College, Columbia University, and New York University. We trained 4 male and 5 female wizard candidates as follows. To familiarize them with the custom database query used in both experiments (described below), trainees were given 24 ASR strings with 5 candidate search results from our previous experiment [7, 8], and asked to select which, if any, of the search results matched the ASR. Next, trainees were given a visual and verbal description of the wizard graphical user interface (*GUI*, also described below), and watched the trainer perform as wizard on a sample call. Each trainee then made five test calls during which she could ask questions and talk to the trainer. We chose as wizards those trainees who were most motivated and skilled at the task. Each caller also made five training calls during

which she could question the trainer via chat.

The trainer was in the room with the wizard during data collection, and could communicate with the caller via chat to coordinate breaks between calls and to restart the system if necessary. This facilitated the complex wizard-caller pair scheduling and dealt with unforeseen difficulties. On the rare occasion of a system crash, the call was not preserved.

Before each call, the caller accessed a web page that provided a *scenario* with patron identity (telephone number, name, and address) plus a list of four books randomly selected from the 3,000 titles used to construct the book title grammar. Each book was described by title, author, and catalogue number. The caller was to request, in any order, one book by title, one by author, one by catalogue number, and the fourth by any of those request types. On each call, the caller first identified herself during the *login* for patron identification, and then ordered the four books.

When a caller telephoned, the wizard interacted with her through two similarly-organized GUIs, one for the login and the other (in Figure 1) for the book requests, with the ASR output at the upper left. Whenever the wizard requested the next book, this ASR output was cleared. Given the ASR output, the wizard could use any substring of the ASR to search for a book by title, author, or catalogue number against the full database of 71,166 books. (Search results appear in the upper right of Figure 1.) A customized query performed partial matching on the ASR string against the database. It used Ratcliff/Obershelp pattern recognition (*R/O*) to evaluate the similarity of the ASR string to a database book title, author, or catalogue number. The R/O score is the number of matching characters divided by the total number of characters [9]. For example, for the ASR "roll dwell" the three top-candidate titles and their R/O scores were CROMWELL (0.666), COLIN POWELL (0.636), and ROBERT LOWELL (0.608).

Although wizards knew that database returns were displayed in decreasing match order, R/O scores did not appear on the GUI. Five candidate books were returned for searches by title or catalogue number. For author searches, up to three candidates were displayed for up to five matching authors, a maximum of 15 search results.

At center left, the book-request GUI displayed how many books had been ordered in the call thus far, details about them, how many questions the wizard had asked, and how often she had asked the caller to repeat. To speak to the caller, the wizard selected a pre-specified prompt that was then forwarded to the text-to-speech component (and spoken to the caller). At center right the GUI displayed prompts the wizard used to advance the dialog: request the next book (with or without implicit confirmation of the book just ordered), inform the caller that the order is complete, offer the caller an optional summary of the order, or say goodbye. The clock in the upper left changed color after six minutes. Wizards were instructed to complete the current book request at that point if it were almost identified, and then end the call, even if all four books had not yet been ordered.

At the bottom of the GUI were 29 question prompts intended to advance the dialogue when the wizard could not

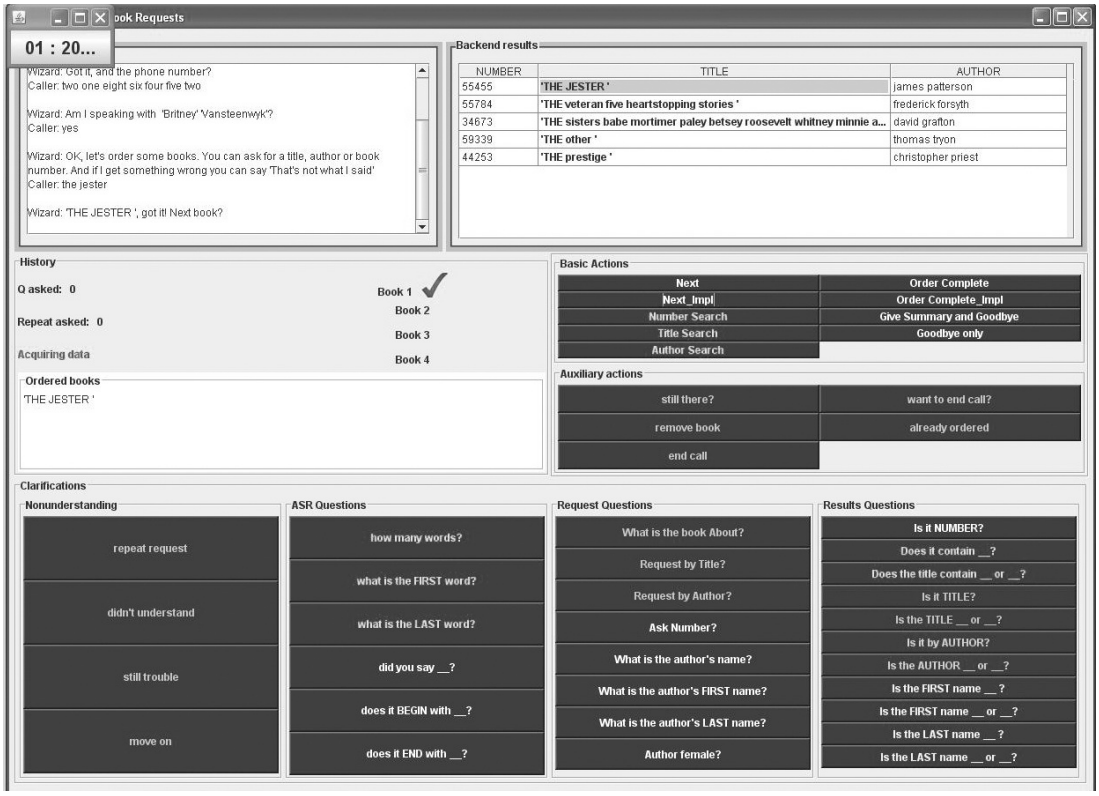

Figure 1: Wizard book-request GUI for ASR "the jester" and a title search.

match a book to the current ASR. Four signaled non-understanding, and asked the caller to repeat or proceed to the next request. Six asked about what the wizard saw in the ASR (e.g., "How many words?"); three allowed the wizard to select one or more words from the ASR to ask about (e.g., "Did you say __?") Eight asked general questions about the book request (e.g., "Did you ask for a book title"), or questions that might elicit a change in *request type* (e.g. "What is the author's name?"). Finally, eleven asked about the search results to disambiguate among the search candidates. These allowed the wizard to make selection from elements of the search results (e.g., "Is the book title __?")

Wizards were surveyed immediately before calls numbered 1, 60, and 120. The first survey collected demographic information. The second and third surveys allowed the wizards to report on their ease with and progress on the task, and elicited strategy information. Callers were surveyed after calls numbered 15, 30, 60, and 90. The survey was always the same; it elicited user satisfaction measures and allowed the callers to make comments.

5. PRELIMINARY DATA ANALYSIS

From 60 wizard-caller pairs (6 wizards and 10 callers) we sought 15 calls per pair, and collected 913 calls. The calls cover 2714 book requests in all, and 20,422 caller utterances. There were 17,288 *adjacency pairs*, portions of dialogue that began with a system prompt and ended with a caller utterance. An adjacency pair contains one or more caller utterances and zero or more database searches. The remainder of this section reports data in the form μ (range, σ) where μ denotes the mean and σ the standard deviation.

On a single call, 2.45 (0 - 5, 1.44) books were ordered, 2.26 (0 - 5, 1.45) of which were correctly identified. (Despite instructions, on two calls 5 books were ordered.) Among all calls, 28% were *fully successful* (all 4 books correctly identified and ordered), and 17% were *failed* (no books correctly identified). Wizards terminated 63% of all calls after the 6-minute time signal. Each call averaged 22.36 (4 - 40, 5.06) caller utterances, with 2.99 (1 - 10, 2.27) words per utterance. Book titles can be long — the average title in the scenarios was 5.96 (1 - 34, 4.38) words. In the full book database, 35% of the titles contain a *subtitle* (an extra phrase that follows the title and is separated from it by a colon). In the random sample of titles used to generate the scenarios for this experiment, 39% contained subtitles. Callers chose whether or not to speak each subtitle.

After each caller utterance, a wizard could ask a question or query the database. Among all adjacency pairs, 32% contained at least one database query. When uncertain about the search results, wizards sometimes attempted multiple queries, on different ASR substrings or with different search types. They averaged 1.09 (1 - 6, 0.33) queries per adjacency pair. Wizards often searched on multiple ASR substrings: 2.9 (1 - 9, 1.76) substrings when searching by title, 2.16 (1 - 8, 1.32) by author, and 2.07 (1 - 8, 1.13) by catalog number. Wizards asked 3.41 (0 - 9, 2.49) questions per book request; only 1% of questions came before any database query at all. Given the ASR string, the wizard chose to search by title, author or number. Of all searches, 43% were by title, 31% by author and 26% by catalogue number. In 28% of the title searches, the correct title appeared among the search results. Author and catalogue number searches returned the correct book 33% and 58% of the time, respectively. When the correct book appeared among the search results, 85% were first on the list, 8 % second, 3% third, and 4% further down the list.

When uncertain about the ASR or book results, wizards selected a question. Wizards could ask for explicit confirmation of a full concept (e.g., ask the caller to confirm the title with a yes/no answer) or of part of a concept (e.g., asked the caller to confirm a single word with a yes/no answer), or confirm implicitly (e.g., have the text-to-speech module speak the title and then ask for the next book). Table 1 reports wizards' question distributions.

5.1 Wizards

Two of the six wizards, WA and WB, most accurately identified the correct books (2.69 and 2.54 correct books per call, respectively; a paired t-test indicates no significant difference). WA is female and WB is male. They also had the fewest failed calls among all the wizards (7% and 11%). Although both were successful, they displayed very different approaches to their task. There are presumably many reasons for this difference. WA is a Masters student and WB an undergraduate; WA majored in linguistics as an undergraduate and WB studies computer science. It is also consistent with the differences in female and male styles of verbal communication noted in the sociolinguistic literature [18].

WA focused on communication, and worked hard to understand the caller's words. She asked more questions per book request than any other wizard (4.09 versus 3.41 for all wizards) and made more database searches per book request than other wizards (2.1 versus 1.77 for all wizards). Among all wizards, WA used the *move-on strategy* (give up on the current book request by asking the caller for the next book) the least often: 0.39 times per call (0.67 for all wizards).

In contrast, WB focused more on the task. He asked questions the least often (2.28 questions per book request). Although he did make several searches to disambiguate the noisy ASR (1.73 database searches per book request), WB also used the move-on strategy more than any other wizard (1.19 times per book request). WA and WB asked similar kinds of questions (Table 1). Most of them concerned the search results or signaled non-understanding. They asked fewer general questions and the fewest questions about the ASR output. WB was the most confident wizard, with the fewest explicit confirmations per call on average. When uncertain, WB preferred to confirm implicitly, and recorded the second most implicit confirmations per call. His task-oriented approach was successful, but sometimes confused

Table 1: The distribution of questions among the four question categories available to the wizards, the average number of confirmations per call, and the average number of questions wizards asked before making any database search.

	All wizards	WA	WB	WE	WD
Questions signaling non-understanding	4334 (37%)	789 (34%)	645 (42%)	613 (33%)	800 (40%)
Questions about the ASR string	788 (8%)	46 (2%)	0 (0%)	293 (15%)	241 (12%)
Questions about the search results	4196 (36%)	854 (36%)	628 (40%)	594 (32%)	529 (26%)
General questions	2244 (19%)	632 (27%)	267 (18%)	368 (20%)	443 (22%)
Average number of explicit confirmations per call	6.07	6.76	4.18	6.59	5.32
Average number of implicit confirmations per call	0.40	0.62	0.68	0.86	0.20
Average number of questions before search	0.35	0.29	0.21	0.42	0.67

the callers (as two callers indicated in the survey). In contrast , and consistent with her communicative approach, WA often asked for confirmation. She had the second-highest rate of explicit confirmations per call, and third highest for implicit confirmations.

The two least successful wizards, WE and WD, had the fewest fully successful calls (16% and 24%). WE had the most failed calls (24%), and both had the fewest correct titles per calls (1.9 and 2.05). WE and WD focused on understanding the ASR without the help of voice search. They asked the most ASR questions, and recorded the most questions per request before any database search (Table 1). WE also made the fewest database queries per adjacency pair on average (1.04 versus 1.09 for all wizards).

5.2 Callers

The caller population was deliberately varied to provide the wizards with a range of recognition difficulties. The best caller, C1, had 3.26 correctly identified books per call on average. 63% of his calls were fully successful and only 6% failed. In contrast, the two worst callers, C0 and C2, averaged 0.96 and 1.13 correct titles per call, respectively. C0 and C2 had only 3% and 6% fully successful calls, and 41% and 43% failed calls, respectively. C1 is male; C0 and C2 are female. All three are native speakers of English. Demographic data collected prior to the experiment indicated that C1 is age 18 – 25; C0 and C2 are age 25 – 35. C0 and C1 have an Eastern seaboard regional accent; C2 has a very slight Indian English accent. All three have a relatively fluent speech quality, although C0's speech rate is slow.

Speech from C1 had the best recognition across request types. Whether wizards searched on title, author, or catalogue number, C1 had the highest percentage of database returns that included the correct book (42%, 55%, and 77%, respectively). The book C1 requested was often returned by the first query; he required the fewest database queries per adjacency pair on average (1.05 versus 1.09 for all callers). C1's well-recognized speech also produced the shortest calls (19.29 utterances and 270.3 seconds per call on average, compared to 22.36 utterances and 345.55 seconds for all callers).

In comparison, speech from C0 had the worst recognition among all callers (only 35% on a catalog number search returned the correct book versus 58% over all callers). Speech from C2 had the worst recognition for titles (only 11% of returns include ed the correct one versus 28% over all callers) and authors (18% versus 33% over all callers). C0 also had the most utterances per calls (23.97 versus 22.36 over all callers). Caller performance was not correlated with utterance length, however. C1 had the third fewest words per utterance (2.82, versus 3 for all callers), while C2 had the third highest (3.11) and C0 the fifth highest (2.98).

Across all callers, catalogue number queries were generally more successful than requests by author or title: the correct book appeared in the return 58%, 33%, and 28% of the time, respectively. C1 not only had the highest percentage of correctly identified books across request type, but also preferred the most recognized query type. Speech from C1 evoked the highest percentage of queries by catalogue number (41% versus 26% for all other callers), and the fewest database queries for title and author (32% and 27% versus 43% and 31% for all other callers). In contrast, speech from C0 evoked the most queries by author (37%). The recognition distribution, however, was not uniform across callers. C3's title and author searches were equally successful (30%). Caller C4 was also atypical. Her title searches returned more correct titles than did her author searches (38% and 30%). These differences among callers also emerged in the caller surveys. C3 reported that the system had difficulty recognizing catalogue numbers, and was better with titles and authors, while C9 reported that the system recognized author names poorly and often mispronounced them.

6. CONCLUSIONS AND FUTURE WORK

In a wizard study of dialogues for book ordering, two differing wizard strategies achieved the greatest success. Our wizards used voice search to contextualize and disambiguate noisy ASR. Some wizards were more confident in their own assessment of ASR accuracy and voice search results, while others asked more questions and confirmed more often. Wizards who relied less on voice search context to disambiguate noisy ASR and asked more questions before making any database query were less successful.

Data from other wizard studies has been used or intended for use to train statistical models of wizard actions [12, 20]. Our earlier experiment demonstrated that we could learn models of wizard behavior with system features. Our newly collected corpus is a rich resource of diverse but successful

wizard behavior, and can be used to train models of that behavior for SDSs. Our wizards' strategies can handle high WER by reference to finer-grained representations, such as using context or phonetic similarity to disambiguate, and to exploit partial recognition. Moreover, competing strategies, such as those modeled on WA's and WB's behavior, could both be implemented in an adaptive system that gauges the best strategy to apply to different users, depending on user preference.

Our corpus, which we will release at the end of our study, is distinguished by its richness. Another corpus, with simulated ASR, had 1,772 turns and 17,076 words [12], compared to our 20,415 user turns and 8,433 words. A different corpus that simulated ASR with a procedure modeled more directly on recognition output included only 144 dialogues compared to our 913 [18]. Our corpus is also distinguished by its collection of 117 runtime features from PocketSphinx, the Phoenix parser, the Helios confidence annotator, the backend and the dialogue history. We expect to extract additional features in post-processing. Previous work on learning dialogue strategies from corpora used much smaller sets of features; 10 features in a study to learn early error detection [16] and 17 features in a study to learn multimodal clarification strategies [13]. Another study to learn non-understanding recovery strategies used approximately 80 features without any feature selection [3]. To our knowledge, there has been no exploration of the kinds of features that best predict different wizard actions.

Given our rich corpus and large set of system features extracted from different dialogue components, our next step is to train models to predict wizard actions with feature selection methods customized for SDSs. We expect that different feature combinations will be best suited to the prediction of different wizard actions, and that feature selection informed by SDS components will support learning the best models. The learned models will be tested in one or more SDSs. Finally, the learned models and particularly relevant features will provide decision rationales, as part of a repertoire of possibly competing strategies such those modeled on WA and WB, for a new SDS architecture currently under construction [5].

This research was supported in part by the National Science Foundation under awards IIS-084966, IIS-0745369, and IIS-0744904.

11. REFERENCES

[1] Bangalore, S., P. Boullier, A. Nasr, O. Rambow, and B. Sagot, "MICA: a probabilistic dependency parser based on tree insertion grammars application note". *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North America Chapter of the Association for Computational Linguistics*, 185-188. Boulder, Colorado, 2009.

[2] Bohus, D. and A.I. Rudniky, *Integrating multiple knowledge sources for utterance-level confidence annotation in the CMU Communicator spoken dialog system.* Technical Report No. CS-190. Carnegie Mellon University, 2002.

[3] Bohus, D. and A.I. Rudniky, "A Principled Approach for Rejection and Threshold Optimization in Spoken Dialogue Systems". *Interspeech 2005*, Lisbon, Spain, 2005.

[4] Bohus, D. and A.I. Rudniky, "The RavenClaw dialog management framework: Architecture and systems.". *Computer Speech and Language*, p. 332-361, 2009.

[5] Epstein, S.L., J.B. Gordon, R.J. Passonneau, and T. Ligorio, "Toward spoken dialogue as mutual agreement.". *To appear in Proceedings of the AAAI-10 Workshop on Metacognition for Robust Social Systems*, Atlanta, Georgia, 2010.

[6] Litman, D., J. Hirschberg, and M. Swerts, "Characterizing and Predicting Corrections, in Spoken Dialogue Systems". *Computational Linguistics*, p. 417-438, 2006.

[7] Passonneau, R., S.L. Epstein, T. Ligorio, and J.B. Gordon, "Learning about Voice Search for Spoken Dialogue Systems". *Proceedings of NAACL-HLT 2010*, Los Angeles, CA, In Press.

[8] Passonneau, R.J., S.L. Epstein, J.B. Gordon, and T. Ligorio, "Seeing what you said: How wizards use voice search results.". *Proceedings of the 6th Workshop on Knowledge and Reasoning in Practical Dialogue Systems, International Joint Conference of Artificial Intelligence*, Pasadena, CA, 2009.

[9] Ratcliff, J.W. and D. Metzener, "Pattern Matching: The Gestalt Approach". *Dr. Dobb's Journal*, p. 46, 1988.

[10] Raux, A. and M. Eskenazi, "A Multi-layer architecture for semi-synchronous event-driven dialogue management.". *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2007)*, Kyoto, Japan, 2007.

[11] Raux, A., B. Langner, A.W. Black, and M. Eskenazi, "Let's Go Public! Taking a spoken dialog system to the real world.". *Interspeech 2005 (Eurospeech)*, Lisbon, Portugal, 2005.

[12] Rieser, V., I. Kruijff-Korvayova, and O. Lemon, "A corpus collection and annotation framework for learning multimodal clarification strategies". *Proceedings of the 6th SIGdial Workshop* Lisbon, Spain, 2005.

[13] Rieser, V. and O. Lemon, "Using Machine Learning to Explore Human Multimodal Clarification Strategies". *COLING/ACL-06*, 659-666. Sidney, Australia, 2006.

[14] Sherwani, J., D. Yu, T. Paek, M. Czerwinski, and A. Acero, "VoicePedia: Towards speech-based access to unstructured information.". *Interspeech 2007*, Antwerp, Belgium, 2007.

[15] Skantze, G., "Exploring Human Error Handling Strategies: Implications for Spoken Dialogue Systems". *Proceedings of ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, Chateau-d'Oex-Vaud Switzerland, 2003.

[16] Skantze, G. and J. Edlund, "Early error detection on word level". *ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction*, Munich, Germany, 2004.

[17] Song, Y., Y. Wang, Y. Ju, M. Seltzer, I. Tashev, and A. Acero, "Voice search of structured media data". *ICASSP 2009*, Taipei, Taiwan, 2009.

[18] Tannen, D., "Gender differences in conversational coherence: Physical alignment and topical cohesion.". *Conversational Coherence and its Development*, p. 167-206, 1990.

[19] Ward, W. and S. Issar, "Recent improvements in the CMU spoken language understanding system.". *ARPA Human Language Technology Workshop*, Plainsboro, NJ, 1994.

[20] Williams, J.D. and S. Young, "Characterizing Task-Oriented Dialogue using a Simulated ASR Channel". *INTERSPEECH 2004 - ICSLP*, Jeju Island, Korea, 2004.