Comparing Spoken Dialog Corpora Collected with Recruited Subjects versus Real Users

Hua Ai[1], **Antoine Raux**[2], **Dan Bohus**[3*], **Maxine Eskenazi**[2], **Diane Litman**[1,4]
[1] Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA 15260, USA
[2] Language Technologies Institute, Carnegie Mellon University, Pittsburgh PA, 15213, USA
[3] Computer Science Department, Carnegie Mellon University, Pittsburgh PA, 15213, USA
[4] Dept. of Computer Science & LRDC, University of Pittsburgh, Pittsburgh, PA 15260, USA
hua@cs.pitt.edu, {antoine,dbohus,max}@cs.cmu.edu, litman@cs.pitt.edu

Abstract

Empirical spoken dialog research often involves the collection and analysis of a dialog corpus. However, it is not well understood whether and how a corpus of dialogs collected using recruited subjects differs from a corpus of dialogs obtained from real users. In this paper we use Let's Go Lab, a platform for experimenting with a deployed spoken dialog bus information system, to address this question. Our first corpus is collected by recruiting subjects to call Let's Go in a standard laboratory setting, while our second corpus consists of calls from real users calling Let's Go during its operating hours. We quantitatively characterize the two collected corpora using previously proposed measures from the spoken dialog literature, then discuss the statistically significant similarities and differences between the two corpora with respect to these measures. For example, we find that recruited subjects talk more and speak faster, while real users ask for more help and more frequently interrupt the system. In contrast, we find no difference with respect to dialog structure.

1 Introduction

Empirical approaches have been widely used in the area of spoken dialog systems, and typically involve the collection and use of dialog corpora. For example, data obtained from human users during Wizard-of-Oz experiments (Okamoto et al., 2001), or from interactions with early system prototypes, are often used to better design system functionalities. Once obtained, such corpora are often then used in machine learning approaches to tasks such as dialog strategy optimization (e.g. (Lemon et al., 2006)), or user simulation (e.g. (Schatzmann et al., 2005)). During system evaluation, user satisfaction surveys are often carried out with humans after interacting with a system (Hone and Graham, 2000); given a dialog corpus obtained from such interactions, evaluation frameworks such as PARADISE (Walker et al., 2000) can then be used to predict user satisfaction from measures that can be directly computed from the corpus.

Experiments with *recruited subjects* (hereafter referred to as *subjects*) have often provided dialog corpora for such system design and evaluation purposes. However, it is not well understood whether and how a corpus of dialogs collected using subjects differs from a corpus of dialogs obtained from *real users* (hereafter referred to as *users*). Selecting a small group of subjects to represent a target population of users can be viewed as statistical sampling from an entire population of users. Thus, (1) a certain amount of data is needed to draw statistically reliable conclusions, and (2) subjects should be randomly chosen from the total population of target users in order to obtain unbiased results. While we believe that most spoken dialog subject experiments have addressed the first point, the second point has been less well addressed. Most academic and many industrial studies recruit subjects from nearby resources, such as college students and colleagues, who are not necessarily representative of the target

[*] Currently at Microsoft Research, Redmond, WA, USA

Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue, pages 124–131,
Antwerp, September 2007. ©2007 Association for Computational Linguistics

users of the final system; the cost to employ market survey companies to obtain a better representation of the target user population is usually beyond the budget of most research projects. In addition, because subjects have either volunteered or are compensated to participate in an experiment, their motivation is often different from that of users. In fact, a recent study comparing spoken dialog data obtained in usability testing versus in real system usage, found significant differences across conditions (e.g., the proportion of dialogs with repeat requests was much lower during real usage) (Turunen et al., 2006).

Our long term goal is to understand the differences that occur in corpora collected from subjects versus users, and to see, if indeed such differences do exist, their impact on empirical dialog research. In this paper we take a first step towards this goal, by collecting and comparing subject and user dialogs with the Let's Go bus information system (Raux et al., 2005). In future work, we plan to investigate how differences found in this paper impact the utility of using subject corpora for tasks such as building user simulations to optimize dialog strategies.

Because there are no well-established standards regarding best practices for spoken dialog experiments with subjects, we first surveyed recent approaches to collecting corpora in laboratory settings. We then used these findings to collect our subject corpus using a "standard" laboratory setting, by adopting the practices we observed in a majority of the surveyed studies. To obtain our user corpus, we collected all dialogs to Let's Go during its deployed hours, over a four day period. Once collected, we quantitatively characterized the two collected corpora using previously proposed measures from the spoken dialog literature. Our results reveal both similarities and differences between the two corpora. For example, we find that while subjects talk more and speak faster, users more frequently ask for help and interrupt the system. In contrast, the dialogs of subjects and users exhibit similar dialog structures.

In Section 2, we describe the papers we surveyed, and summarize the common practices we observed for collecting dialog corpora using subjects. In Section 3, we introduce the Let's Go spoken dialog system, which we use to collect both our subject and user corpora. In Section 4, we describe the specific in-lab experiment we conducted with recruited subjects. We then introduce the evaluation measures used for our corpora comparisons in Section 5, followed by a presentation of our results in Section 6. Finally, we further discuss and summarize our results in Section 7.

2 Literature Review

In this section we survey a set of spoken dialog papers involving human subject experiments (namely, (Allen et al., 1996), (Batliner et al., 2003), (Bohus and Rudnicky, 2006), (Giorgino et al., 2004), (Gruenstein et al., 2006), (Hof et al., 2006), (Lemon et al., 2006), (Litman and Pan, 2002), (Möller et al., 2006), (Rieser et al., 2005), (Roque et al., 2006), (Singh et al., 2000), (Tomko and Rosenfeld, 2006), (Walker et al., 2001), (Walker et al., 2000)), in order to define a "standard" laboratory setting for use in our own experiments with subjects. We survey the literature from four perspectives: subject recruitment, experimental environment, task design, and experimental policies.

Subject Recruitment. Recruiting subjects involves deciding who to recruit, where to recruit, and how many subjects to recruit. In the studies we surveyed, the number of subjects recruited for each experiment ranged from 10 to 72. Most of the studies recruited only native speakers. Half of the studies clearly stated that the subjects were balanced for gender. Most of the studies recruited either college students or colleagues who were not involved in the project itself. Only one study recruited potential system users by consulting a market research company.

Experimental Environment. Setting up an experimental environment involves deciding where to carry out the experiment, and how to set up this experimental environment. The location of the experiment may impact user performance since people behave differently in different environments. This factor is especially important for spoken dialog systems, since system performance is often impacted by noisy conditions and the quality of the communication channel. Although users may call a telephone-based dialog system from a noisy environment using a poor communication channel (e.g., by using a cell phone to call the system from the street), most experiments have been conducted in a quiet in-room lab setting. Subjects typically talk to the system directly

via a high-quality microphone, or call the system using a land-line phone. Among the studies we looked at, only 2 studies had subjects call from outside the lab; another 2 studies used driving simulators. One study changed the furniture arrangement in the lab to simulate home versus office scenarios.

Task Design. Task design involves specifying whether subjects should use the dialog system to accomplish specific tasks, and if so, defining those tasks. All except one study asked subjects to finish a set of fixed tasks in a predefined order. In one study, subjects were asked to do 2 open tasks after a series of 7 fixed tasks. In another study, where the system provided restaurant information, the researchers asked the subjects to ask about information for at least 4 restaurants, but did not specify the restaurant names. The number of tasks in these studies ranged from 2 to 10.

Experimental Policies. Experimental policies involves specifying additional procedures for running subjects during the course of the experiment. None of the studies mentioned that they controlled their experiments by setting any time limits for the subjects. Only 2 studies clearly declared that subjects were told to read some instructions before the experiment started. While two studies motivated subjects by offering a bonus upon task completion, the majority of studies paid subjects on the basis of their participation alone.

In summary, a standard way to carry out human subject experiments with spoken dialog systems (where we use standard to mean that the practice occurred in a majority of the papers surveyed), is as follows: (1) Recruit at least 10 subjects who are college students or colleagues who are native English speakers, trying to balance between genders; (2) Ask the subjects to come to the lab to generate their dialogs with the system; (3) Set up several tasks for the subjects, and ask them to complete these tasks in a certain order; (4) Pay the subjects for their participation, without a bonus. As will be seen in Section 4, we follow these practices when designing our own experiment.

3 System Description

The study described in this paper was conducted in the Let's Go Lab which uses the Let's Go bus information system, a telephone-based dialog system that provides schedule information for buses in the Pittsburgh area (Raux et al., 2005). The Lab is a service run by the creators of Let's Go to allow other researchers access to their numerous users to run experiments. When the customer service line of the Port Authority of Allegheny County (which manages buses in Pittsburgh) is not staffed by operators (i.e. from 7pm to 6am on weekdays and 6pm to 8am on weekends), callers are redirected to Let's Go. In the Let's Go Lab, experimenters typically run offline and/or in-lab experiments first, then evaluate their approach using the live system.

S: Welcome to the CMU Let's Go bus information system. To get help at any time, just say Help or press zero. What can I do for you?
U: UH I NEED TO GET FROM FIFTH AND BIGELOW
S: Leaving from Fifth and Bigelow . Did I get that right?
U: YES
S: Right. What is your destination?
U: AIRPORT
S: Going to airport . Did I get that right?
U: YES
S: Okay. When are you going to take that bus?
U: NINE P UH... SEVEN P M
S: Leaving at 7 p.m. . Is this correct?
U: YES
S: Right. Just a minute. Let me check that for you. There is a 28X leaving Fifth Avenue at Bigelow . . .

Figure 1: Example Dialog with Let's Go.

An example dialog with Let's Go (obtained from a subject) is shown in Figure 1. The interaction with the system itself starts with an open prompt ("What can I do for you?") followed by a more directed phase where the system attempts to obtain the missing information (origin, destination, travel time, and optionally route number) from the user. Finally, the system provides the best matching bus number and time, at which point the user has the possibility of asking for the next/previous buses.

Let's Go is based on the Olympus architecture developed at CMU (Bohus et al., 2007). It uses the RavenClaw dialog manager (Bohus and Rudnicky, 2003), the PocketSphinx speech recognition

High-level dialog features	
number of turns	turn
duration of dialog	dialogLen
total words per user turn	U_word
number of dialog acts per system/user turn	U_action, S_action
ratio of system and user actions	Ratio_action
Dialog style/cooperativeness	
dialog acts	S_requestinfo, S_confirm, S_inform, S_other, U_provideinfo, U_yesno, U_unknown
Task success/efficiency	
average goal/subgoal achievement rate	success%
Speech recognition quality	
non-understanding rate	rejection%
average ASR confidence score	confScore
User dialog behavior	
requests for help	help%
touch-tone	dtmf%
barge-in	bargein%
speaking rate	speechRate

Figure 2: Evaluation Measures (and abbreviations).

engine (Huggins-Daines et al., 2006) and a domain-specific voice built with the Festival/Festvox toolkit (Black and Lenzo, 2000) and deployed on the Cepstral Swift engine (Cepstral, LLC, 2005). As of April 2007, the system has received more than 34,000 calls from the general public, all of which are recorded with logs and available for research.

4 Experimental Setup

Our experiment involves collecting, then comparing, two types of dialog corpora involving human users and Let's Go. Here we describe how we collected our *subject corpus* and our *user corpus*, i.e., our two experimental conditions. The same version of Let's Go was used by the users and the subjects.

To collect our subject corpus we used a "standard" laboratory experiment, following typical practices in the field as summarized in Section 2. We recruited 39 subjects (19 female and 20 male) from the University of Pittsburgh who were native speakers of American English. We asked the subjects to come into our lab to call the system from a land-line phone. We designed 3 task scenarios[1] and asked the subjects to complete them in a given sequence. Each task included a departure place, a destination, and a time restriction (e.g., going from the University of Pittsburgh to Downtown, arriving before 7PM). We used map representations of the places and graphic representations of the time restrictions to avoid influencing subjects' language. Subjects were instructed to make separate calls for each of the 3 tasks. As shown in Figure 1, the initial system prompt informed the users that they could say "Help" at any time. We did not give any additional instructions to the subjects on how to talk to the system. Instead, we let the subjects interact with the system for 2 minutes before the experiment, to get a sense of how to use the system. Subjects were compensated for their time at the end of the experiment, with no bonus for task completion. Although we set a time limit of 15 minutes as the maximum time per task, none of the subjects reached this limit.

For our user corpus, we used 4 days of calls to Let's Go (two days randomly chosen from the weekday hours of deployment, and two from the weekend hours of deployment) from the general public. Recall that during nights and weekends, callers to the Port Authority's customer service line are redirected to Let's Go.

5 Evaluation Measures

To examine whether differences exist between our two corpora, we will use the evaluation measures shown in Figure 2. All of these measures are adopted from prior work in the dialog literature.

Schatzmann et al. (2005) proposed a comprehensive set of quantitative evaluation measures to compare two dialog corpora, divided into the following three types: high-level dialog features, dialog style/cooperativeness, and task success/efficiency.

[1] It should be noted that one of these tasks required transferring to another bus, which was not explicitly handled by the system. This task was therefore particularly difficult to complete, especially for subjects not familiar with the Port Authority network. However, because this task represented a situation that users might face, we still included this task in the study.

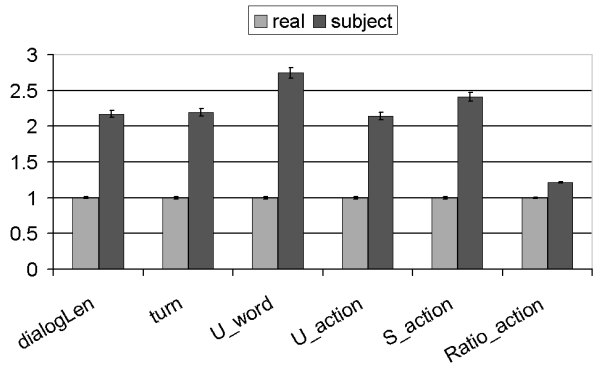

Figure 3: Comparing High-level Dialog Features.

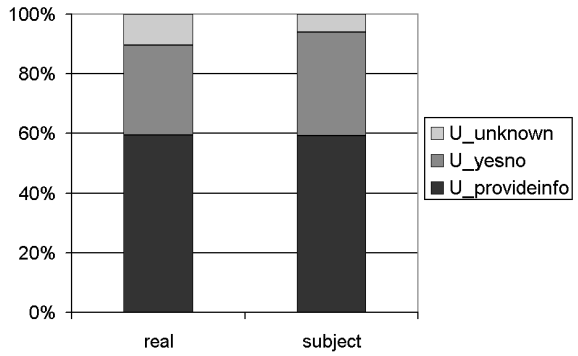

Figure 4: Comparing User Dialog Acts.

We adapt these measures for use in our comparisons, based on the information available in our corpora. For high-level dialog features (which capture the amount of information exchanged in the dialog) and dialog style, we define and count a set of system/user dialog acts. On the system side, **S_requestinfo**, **S_confirm**, and **S_inform** indicate actions through which the system respectively requests, confirms, or provides information. **S_other** stands for other types of system prompts. On the user side, **U_provideinfo** and **U_yesno** respectively identify actions by which the user provides information and gives a yes/no answer, while **U_unkown** represents all other user actions. Finally, **S_action** (resp. **U_action**) represents any of the system (resp. user) actions defined above, and **Ratio_action** is the ratio between **S_action** and **U_action**.

We also define a variety of other measures based on other studies (e.g., (Walker et al., 2000; Turunen et al., 2006)). Two of our measures capture speech recognition quality: the non-understanding rate (**rejection%**) and the average confidence score (**confScore**). In addition, we look into how frequently the users ask for help (**help%**), how often they use touchtone (**dtmf%**), how often they interrupt the system (**bargein%**), and how fast they speak (**speechRate**, number of words per second).

All of the features used to compute our evaluation measures are automatically extracted from system logs. Thus, the user dialog acts and dialog behavior measures are identified based on speech recognition results. For **success%**, we consider a task to be completed if and only if the system is able to get enough information from the user to start a database query and inform the user of the result (i.e., either specific bus schedule information, or a message that the queried bus route is not covered by the system).

6 Results

Our subject corpus consists of 102[2] dialogs, while our user corpus consists of 200 dialogs (90 obtained during 2 weekdays, and 110 obtained over a weekend). To compare these two corpora, we compute the mean value for each corpus with respect to each of the evaluation measures shown in Figure 2. We then use two-tailed t-tests to compare the means across the two corpora. All differences reported as statistically significant have p-values less than 0.05 after Bonferroni corrections.

As a sanity check we first compared the weekday and weekend parts of the user corpus with respect to our set of evaluation measures. None of the measures showed statistically significant differences between these two subcorpora.

Figure 3 graphically compares the means of our high-level dialog features, for both the user and subject dialog corpora. In the figures, the mean values of each measure are scaled according to the mean values of the user corpus, in order to present all of the results on one graph. For example, to plot the means of **dialogLen**, we treat the mean **dialogLen** of the user corpus as 1 and divide the mean **dialogLen** of the subject corpus by the mean of the user corpus. The error bars show the standard er-

[2]Some subjects mistakenly completed more than one task per dialog. Such multi-task dialogs were not included in our analysis, because our evaluation measures are calculated on a per-dialog basis

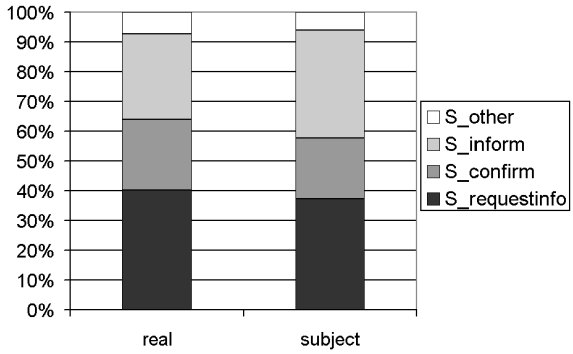

Figure 5: Comparing System Dialog Acts.

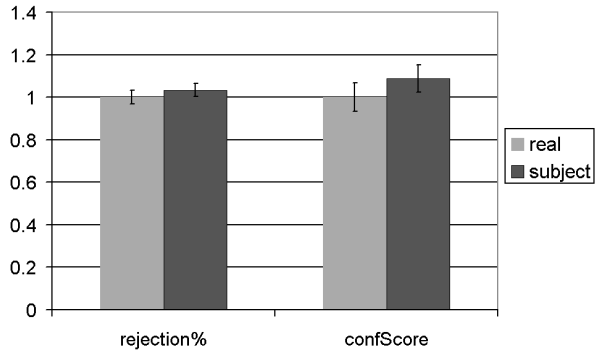

Figure 6: Comparing Speech Recognition Quality.

rors. Using t-tests on the unnormalized means (described above), we confirm that the user dialogs and the subject dialogs are significantly different on all of the high-level dialog features. Subjects talk significantly more than users in terms of number of words per utterance; the number of turns per dialog is also higher for subjects. **U_action** and **S_action** show that both the system and the user transmit more information in the subject dialogs. **Ratio_action** shows that subjects are more passive than users, in the sense that they produce relatively less actions than the system.

Figure 4 (resp. Figure 5) shows the distribution of the user (resp. system) actions in both the user and subject corpora. Subjects give more yes/no answers and produce fewer unrecognized actions than users (these differences are statistically significant). On the other hand, there is no significant difference in **U_provideinfo** between users and subjects. The system provides significantly more information (**S_inform**) to the subjects than to the users, which is consistent with the fact that the task completion rate is higher for subjects. Using automatic indicators to estimate task completion as discussed in Section 5, we find that the completion rate for subjects is 80.7%, while for users it is only 67%. There are also significantly more **S_other** in dialogs with users than with subjects. We did not find any significant difference in the number of system requests (**S_requestinfo**) or confirmations (**S_confirm**).

Figure 6 shows the results for speech recognition quality, using scaled mean values as in Figure 3. There are no statistically significant differences between the number of rejected user turns or the average confidence scores of the speech recognizer. Recall, however, that these measures are automatically calculated using recognition results. Until we can examine speech recognition quality using manual transcriptions, we believe that it is premature to conclude that our speech recognizer performs equally well in real and lab environments.

Figure 7 shows the normalized mean values and standard errors for our user dialog behaviors. Our results agree with the findings in (Turunen et al., 2006). All four measures show significant differences between user and subject dialogs. Users barge in more frequently, use more DTMF inputs, and ask for more help than subjects, while subjects speak faster than users.

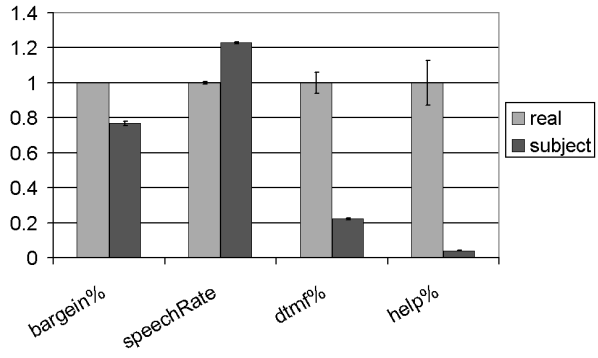

Figure 7: Comparing User Dialog Behaviors.

To summarize, subject dialogs are longer and contain more caller actions than user dialogs, suggesting that subjects are more patient and try harder than users to complete their tasks. In addition, there are less barge-ins and unknown dialog acts in sub-

ject dialogs. Subjects also appear to speak faster than users. This may be because subjects are calling the system in very controlled and quiet conditions, whereas users may experience a higher cognitive load due to their environment (e.g. calling from the street) or emotional state (e.g. concerned about missing a bus).

Finally, in addition to comparing our corpora on the dialog level, we also present a brief examination of the differences between the first user utterances from the dialogs in each corpus. (Because we are only looking at a small percentage of our user utterances, here we are able to use manual transcriptions rather than speech recognition output.) The impact of open system initial prompts on user initial utterances is an interesting question in dialog research (Raux et al., 2006). Most users answer the initial open prompt of Let's Go ("What can I do for you?") with a specific bus route number, while subjects often start with a departure place or destination. Subject queries may be restricted by the assigned task scenarios. However, it is interesting to note that many users call the system to obtain schedule information for a bus route they already know, rather than to get information on how to reach a destination. We also observe that there are only 2% void utterances (when only background noise is heard) in subject dialogs, while there are 20% in user dialogs. This confirms that subjects and users dialog with the system in very different environments.

7 Conclusions and Discussion

In this paper, we investigated the differences between dialogs collected with users in real settings and with subjects in a standard lab setting, and observed statistically significant differences with respect to a set of well-known dialog evaluation measures. Specifically, our results show that subjects talk more with the system and speak faster, while users barge in more frequently, use more touchtone input and ask for more help. Although there are some significant differences in the frequency of particular system/user dialog acts, there is no significant difference in the overall ratios of different dialog acts (i.e., the structure of the dialogs is similar).

Many of the differences we observed suggest that, because users and subjects have different behaviors, a system that is optimal for one population might not be for the other. For instance, the fact that users resort more to system help than subjects and at the same time barge in more often implies different designs for help prompts. Such prompts should be shorter for users to avoid information overload (and early barge-in which prevents them from hearing the message), but might include more information for subjects.

Our results also offer insights for user simulation training. Most current research simulates user behavior on the dialog act level. In this case, training the simulation models from a user corpus or from a subject corpus may not differ much since the dialog act distributions were shown to be similar in our two corpora. At the speech/word level, however, we did see significant differences in user behavior. Thus, simulations trained on subject corpora may be insufficient to train systems that explore problems such as barge-in, switch between modalities, and so on.

Finally, our work can contribute to an understanding of how Let's Go Lab can satisfy the needs of the spoken dialog community. By charting the differences between users and subjects, we can determine how tests carried out on the Lab can translate back to the academic systems of the experimenters.

Acknowledgments

This work is supported by the US National Science Foundation under grants number 0208835 and 0325054. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. We would like to thank the Port Authority of Allegheny County for their help in making the Let's Go system accessible to Pittsburghers.

References

J. F. Allen, B. W. Miller, E. K. Ringger, and T. Sikorski. 1996. *A Robust System for Natural Spoken Dialogue*. In Proceedings of the 34nd Annual Meeting of the Association for Computational Linguistics (ACL).

A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Noth. 2003. *How to Find Trouble in Communication*. Speech Communication, Vol. 40, No. 1-2, pp. 117-143.

A. W. Black and K. Lenzo. 2000. *Building Voices in the Festival Speech System*. http://festvox.org/bsv/

D. Bohus and A. Rudnicky. 2003. *RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda*. In Proceedings of Eurospeech 2003, Geneva, Switzerland.

D. Bohus and A. Rudnicky. 2006. *A K Hypotheses + Other Belief Updating Model*. In AAAI Workshop on Statistical and Empirical Approaches to Spoken Dialogue Systems.

D. Bohus, A. Raux, T. K. Harris, M. Eskenazi, and A. Rudnicky. 2007. *Olympus: an open-source framework for conversational spoken language interface research*. In Proceedings of the HLT-NAACL 2007 workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology, Rochester, NY, USA.

Cepstral, LLC. 2005. *SwiftTM: Small Footprint Text-to-Speech Synthesizer*. http://www.cepstral.com

D. Huggins-Daines, M. Kumar, A. Chan, A. W Black, M. Ravishankar, and A. I. Rudnicky. 2006. *Pocket-Sphinx: A Free, Real-Time Continuous Speech Recognition System for Hand-Held Devices*. In Proc. of ICASSP 2006.

T. Giorgino, S. Quaglini, and M. Stefanelli. 2004. *Evaluation and Usage Patterns in the Homey Hypertension Management Dialog System*. Dialog Systems for Health Communication, AAAI Fall Symposium, Technical Report FS-04-04

A. Gruenstein, S. Seneff, and C. Wang. 2006. *Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Databases*. In Proc. of ICSLP, 2006.

A. Hof, E. Hagen and A. Huber. 2006. *Adaptive Help for Speech Dialogue Systems Based on Learning and Forgetting of Speech Commands*. In Proc. of 7th SIGdial.

K. S. Hone and R. Graham. 2000. *Towards a tool for the subjective assessment of speech system interfaces (SASSI)*. Natural Language Engineering, 6(3/4), 287-305.

O. Lemon, K. Georgila, J. Henderson. 2006. *Evaluating Effectiveness and Portability of Reinforcement Learned Dialogue Strategies with real users: the TALK TownInfo Evaluation*. In Proceedings of IEEE/ACL Spoken Language Technology.

D. J. Litman and S. Pan. 2002. *Designing and Evaluating an Adaptive Spoken Dialogue System*. User Modeling and User-Adapted Interaction. Vol. 12, No. 2/3, pp. 111-137

S. Möller, R. Englert, K. Engelbrecht, V. Hafner, A. Jameson, A. Oulasvirta, A. Raake, and N. Reithinger. 2006. *MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations*. In Proc. ICSLP2006.

M. Okamoto, Y. Yang, and T. Ishida. 2001. *Wizard of oz method for learning dialog agents*. Cooperative Information Agents V, volume 2182 of LNAI, pages 20–25.

A. Raux, B. Langner, D. Bohus, A. W Black, M., Eskenazi. 2005. *Let's Go Public! Taking a Spoken Dialog System to the Real World*. In Proceedings of Interspeech 2005 (Eurospeech), Lisbon, Portugal.

A. Raux, D. Bohus, B. Langner, A. W Black, M., Eskenazi. 2006. *Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience*. In Proceedings of Interspeech 2006.

V. Rieser, I. Kruijff-Korbayova, and O. Lemon. 2005. *A corpus collection and annotation framework for learning multimodal clarification strategies*. In Proceedings of SIGdial 2005.

A. Roque, A. Leuski, V. Rangarajan, S. Robinson, A. Vaswani, S. Narayanan, and D. Traum. 2006. *Radiobot-cff: A spoken dialogue system for military training*. In Proceedings of International Conference on Spoken Language Processing 2006.

J. Schatzmann, K. Georgila, and S. Young. 2005. *Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems*. Proceedings of 6th SIGdial Workshop on Discourse and Dialogue.

S. P. Singh, M. J. Kearns, D. J. Litman, and M. A. Walker. 2000. *Empirical Evaluation of a Reinforcement Learning Spoken Dialogue System*. Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence.

S. Tomko and R. Rosenfeld. 2006. *Shaping user input in speech graffiti: a first pass*. CHI Extended Abstracts.

M. Turunen, J. Hakulinen and A. Kainulainen. 2006. *Evaluation of a Spoken Dialogue System with Usability Tests and Long-term Pilot Studies: Similarities and Differences*. In Proceedings of Interspeech 2006.

M. Walker, J. Aberdeen, J. Boland, E. Bratt, J. Garofolo, L. Hirschman, A. Le, S. Lee, S. Narayanan, K. Papineni, B. Pellom, J. Polifroni, A. Potamianos, P. Prabhu, A. Rudnicky, G. Sanders, S. Seneff, D. Stallard, and S. Whittaker. 2001. *DARPA Communicator dialog travel planning systems: The June 2000 data collection*. In Proc. EUROSPEECH.

M. A. Walker, C. A. Kamm, and D. J. Litman. 2000. *Towards Developing General Models of Usability with PARADISE*. In Natural Language Engineering, Vol. 6, No. 3.