Proceedings of the Fourth Annual

# Young Researchers’ Roundtable on Spoken Dialog Systems

Columbus, Ohio
21st – 22nd June 2008

http://www.yrrsds.org

# Sponsors

VoiceObjects

NUANCE

Microsoft Research

Google

at&t

SpeechStorm

IBM Research

# Endorsements

ISCA
ACL 2007
SIGdial

# Organizing Committee

| | |
|---|---|
| Hua Ai | Intelligent Systems Program, University of Pittsburgh, USA |
| Carlos Gómez Gallo | Department of Computer Science, University of Rochester, USA |
| Robert J. Ross | Department of Computer Science, University of Bremen, Germany |
| Sabrina Wilske | Department of Computational Linguistics, Saarland University, Germany |
| Andi Winterboer | Institute for Communicating and Collaborative Systems, University of Edinburgh, UK |
| Craig Wootton | University of Ulster, Belfast, Northern Ireland |

# Local Organization

| | |
|---|---|
| Tim Weale | Department of Computer Science and Engineering, Ohio State University, USA |

# Advisory Committee

| | |
|---|---|
| John Bateman | University of Bremen, Germany |
| Robert Dale | Macquarie University, Australia |
| Sudeep Gandhe | University of Southern California, USA |
| Stefan Hamerich | Harman/Becker Automotive Systems, Germany |
| Hartwig Holzapfel | University of Karlsruhe, Germany |
| Kristiina Jokinen | University of Helsinki, Finland |
| Tatsuya Kawahara | Kyoto University, Japan |
| Alistair Knott | Otago University, New Zealand |
| Geert-Jan Kruijff | DFKI Language Technology, Germany |
| Diane Litman | University of Pittsburgh, USA |
| Michael McTear | University of Ulster, UK |
| Helen Meng | Chinese University of Hong Kong, Hong Kong |
| Johanna Moore | University of Edinburgh, UK |
| Ian O'Neill | Queen's University Belfast, UK |
| Tim Paek | Microsoft Research, USA |
| Verena Rieser | University of Edinburgh, UK |
| Antonio Roque | University of Southern California, Los Angeles, USA |
| David Schlangen | University of Potsdam, Germany |
| Marilyn Walker | University of Sheffield, UK |
| Fuliang Weng | Bosch Research, USA |
| Wieneke Wesseling | University of Amsterdam, Netherlands |
| Michael White | Ohio State University, USA |

# Foreword from the Organizers

Welcome to the Forth annual Young Researchers' Roundtable on Spoken Dialogue Systems. The workshop carries on from previous workshops, held at were held in Antwerp (Interspeech 2007), Pittsburgh (Interspeech 2006) and Lisbon (Interspeech 2005).

The purpose of the Young researchers' Roundtable is to promote networking and discussion in the field of dialogue systems amongst student and young researchers in both academia and industry. It is our hope today that our program will both facilitate and encourage this, and our time together will be worthwhile, productive, and fun! We are pleased that our number of participants has increased on last year to 33 participants, reflecting the growing interest in the field of dialogue research from both parties.

Given the high standards regarding the program of previous roundtables, it was a difficult task for the organizers to further improve this year's schedule; however, we are confident that this will be achieved with the inclusion of some new events to the roundtable. Our drinks reception on Friday night will provide an opportunity for all involved to get to know one another before the roundtable begins. Next we are pleased to see the inclusion of 3 industrial presentations, given by student researchers in industry from VoiceObjects, AT&T and Nuance. Continuing on from the success of last years' academic panel, we have this year once again decided to hold a panel of senior researchers, this time from both academia and industry. We are pleased to announce our panel of Alex Rudnicky, David Schlangen and Tim Paek. Lastly, we are excited also about the first half day event to follow the main roundtable day, with a special session on evaluating spoken dialogue systems. It is our hope that these additional elements will further the past success of the Young Researchers' Roundtable, and make it even more enjoyable for our participants.

This year's roundtable was supported by VoiceObjects, Nuance, Microsoft Research, Google, AT&T Labs-Research, SpeechStorm and IBM Research. We thank these sponsors for giving us the opportunity to make the workshop possible for our participants at a low fee and providing coffee breaks, lunch, and dinner. We also received endorsements from ISCA, ACL 2008, and SIGdial. Thanks to these institutions for promoting the event. Special thanks to the Ohio state University for enabling us to use the university's rooms. Further thanks to the members of our advisory committee for providing assistance and helping to promote the event. Finally we would like to thank this year's participants, who provided interesting position papers and thoughtful questions for discussion.

Wishing you all an interesting day in Columbus and a successful roundtable,
The organizing committee of YRRSDS 2008
*Hua Ai*
*Carlos Gómez Gallo*
*Robert J. Ross*
*Tim Weale*
*Sabrina Wilske*
*Andi Winterboer*
*Craig Wootton*

# Table of Contents

# Workshop Program

**Friday**
Drinks Reception

**Saturday, June 21st**
8:30 Registration / breakfast
9:00 Introductions

9:45 Discussion session I
- Dialog Strategy Learning and dialog design
- Next killer-apps
- Multimodal Systems

11:30 Summaries and discussion

11:45 Industry/Company talks:
- Tobias Göbel (VoiceObjects)
- Jason Williams (AT&T)
- Simona Gandrabur (Nuance)

12:45 Lunch + set up demos/posters

13:45 Discussion session II
- Empirical Approach -- Training from Dialog Corpora
- Realistic conversation -- How to make SDS human-like?
- Dialog System development.

15:30 Summaries and discussion

15:45 Introduction of Industry and Academic Panellists
- Alex Rudnicky (CMU)
- David Schlangen (University of Potsdam)
- Tim Paek (Microsoft)

16:00 Afternoon coffee with poster and demo presentation session

16:45 Industry/Academic Panel Session

18:30 Dinner

**Sunday, June 22nd**
9:30 Breakfast

9: 00 Special session: *Frameworks and Grand Challenges for Dialog System Evaluation*
11:30 Discussion

12:30 Lunch

# Industry Speakers

## Tobias Göbel

Employer: VoiceObjects

Presentation Abstract: In addition to text (SMS) and voice, current mobile devices have built-in capabilities for the display of videos as well as Web pages that allow for new kinds of interactive man-machine applications. The talk describes a framework for building, deploying, and analyzing phone applications that can run in the voice, video, text, and Web channel. It shows that these channels have enough similarities so that applications can be generated by the same dialog definition.

Tobias Göbel studied computational linguistics, phonetics, and computer science at the universities of Bonn (Germany) and Edinburgh (UK). After graduating in 2003 he has held different positions at VoiceObjects, the Phone Application Server company, and worked as a contract teacher at the University of Bonn. As Partner Consultant at VoiceObjects he was involved in VUI design and implementation of various large voice portals. As a Program Manager, he held responsibilities for the development of the VoiceObjects product family, mainly for the VoiceXML-based runtime server. Today he is acting as Senior Presales Consultant for the company.

## Jason Williams

Employer: AT&T

Presentation Abstract: In this brief talk, I'll describe my experience of finishing a PhD and finding a research job 2-3 years ago. I'll share some thoughts about what seemed important then, and what seems important now. I'll also talk a bit about AT&T's research lab, including areas of research, culture, opportunities for internships and staff positions, and the application process. Time permitting, I'll also briefly describe my work at AT&T.

Jason D. Williams is a Principal Member of Technical Staff at AT&T Labs - Research in Florham Park, New Jersey, USA. He received a BSE in Electrical Engineering from Princeton University in 1998, and at Cambridge University he received an M Phil in Computer Speech and Language Processing in 1999 under a Churchill Scholarship and a PhD in Information Engineering in 2006 under a Gates Scholarship. His main research interests are dialog management, the design of spoken language systems, and planning under uncertainty. He has previously held positions at Tellme Networks, Edify Corporation (now Intervoice), and McKinsey & Company's Business Technology Office.

## Simona Gandrabur

Employer: Nuance

Presentation Abstract: There is a well-known gap between academic dialogue research, which tends to explore advanced methods for dynamic, adaptive, data-driven methods for open, unconstrained dialogues, and the current state-of-the art commercial dialogue applications, which are mostly constrained, directed-prompt, typically based on an exhaustive Finite State Machine (FSM) description of the call-flow. Several more-or-less independent efforts have been undertaken to address this issue, such as the "Bridging the Gap: Academic and Industrial Research in Dialog Technologies" workshop at NAACL-HLT 2007 and the VoiceXML Advanced Dialog Working Group. At Nuance, two projects are currently focusing on advanced dialogue technologies that facilitate the development of commercial natural language dialogue applications for embedded or enterprise platforms. The emphasis is on facilitating the packaging, code reuse, and ease-of-use during application development, maintenance and tuning. In this talk I will give an overview of these projects, their motivation and their goals.

Simona has studied mathematics (University of Bucharest) and then Computer Science (Université de Montréal), graduating in 2001 with a PhD in formal specification and verification for hardware interfaces. From 1999 until 2002 she has been part of the speech technology team at Locus Dialogue, working mainly on confidence measures and semantic interpretation and leading the ASR Post-Processing team.

From 2002 until 2005 Simona was a post-doctoral fellow at the RALI computational linguistic lab (Université de Montréal), where she has investigated the use of machine learning techniques for confidence estimation on various statistical natural language processing technologies (s.a. machine translation and language identification). During this time she has also co-lead a 6 week workshop on this topic at the Johns Hopkins University CSLP lab, in the summer of 2003. Following Simona's post-doc, She worked for two years at Idilia, a Montreal start-up specialized in statistical word sense disambiguation, where she lead the Named Entity Recognition team.

Simona joined Nuance Communications in September 2007 within the Enterprise Natural Language Understanding (NLU) Research team, working on advanced dialogue research. Within this project the team is developing a dynamic, adaptive dialogue management infrastructure that handles unconstrained and unsolicited user input in task-oriented speech applications.

# Panelists

## Alex Rudnicky

Affiliation: Carnegie Mellon University

Dr. Rudnicky's research has spanned many aspects of spoken language, including knowledge-based recognition systems, language modeling, architectures for spoken language systems, multi-modal interaction, the design of speech interfaces and the rapid prototyping of speech-to-speech translation systems. His most recent work has been in spoken dialog systems, with contributions to dialog management, language generation and the computation of confidence metrics for recognition and understanding. Dr. Rudnicky has published over 60 refereed papers and is a recipient of the Allen Newell Award for Research Excellence.

Dr. Rudnicky is currently a Principal Systems Scientist in the Computer Science Department at Carnegie Mellon University and on the faculty of the Language Technologies Institute. He serves on the boards of the Applied Voice Input/Output Society (AVIOS) and of SIGdial.

## David Schlangen

Affiliation: University of Potsdam

David studied Computational Linguistics, Computer Science and Philosophy at the Universities of Bonn and Edinburgh, graduating in 1999 from Bonn. Afterwards, he returned to Edinburgh to pursue a PhD (under the supervision of Alex Lascarides), developing a logical model of the interpretation of non-sentential utterances.

In 2003, he joined the Computational Linguistics Group in Potsdam, Germany, as a Post-Doc, working, among other things, on interaction management phenomena in dialogue. In 2006, he was awarded an "Emmy Noether Grant" from DFG, and now leads an Independent Research Group at Potsdam, working on computational models of the timing of linguistic and non-linguistic
behaviours in dialogue.

## Tim Paek

Affiliation: Microsoft

Tim Paek is a researcher at Machine Learning and Applied Statistics Group, Microsoft Research. His research focuses on fostering and improving human interaction with automated systems, and in particular, those capable of engaging in conversational dialogue. He currently serves on the Scientific Advisory Committee of (SIGDIAL), the Special Interest Group on Discourse and Dialogue for ACL and ISCA. He is also on the Editorial Board of the Journal of Dialogue Systems.

**Jaime C. Acosta**

Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968

jcacosta@miners.utep.edu

## 1 Research Interests

My research interest is in making a persuasive dialog system that incorporates **automated interpretation and generation of emotion**. My thesis will investigate which elements from a speech signal should be extracted to model the affective state of the user. The adaptive SDS will utilize the speaker's state to dynamically calibrate the motivator's actions and generate emotional dialog. The generated dialog will have an emphasis on gaining **rapport and trust**, therefore allowing users to feel more comfortable and more accepting of the technology. By accomplishing this, users of automated motivational systems, health-care systems, and automated task-oriented systems in general will benefit. In addition, non-task oriented systems, such as virtual humans in video games will be more realistic.

### 1.1 Previous and Current Work

Previous work towards the persuasive dialog systems has included collecting a *Persuasive Advisor* corpus consisting of ten dialog sessions between a graduate school advisor and an undergraduate student. The dialog acts for both the advisor and the student were labeled and analyzed and a system that will mimic the advisor's dialog is currently being built.

The dialog system will strive to motivate undergraduate students to consider attending graduate school. It will be system-directed and will begin by greeting users and obtaining information about their current state on the subject. Different strategies, which are based on some predefined tasks from the corpus, will be prompted to the listener depending on what seems to be the most effective strategy at the time. Certain strategies will either enable or disable different tasks such as advice about financial, GRE scores, GPA, and others. Enabled tasks will be candidates for future prompts.

This application will be built with VoiceXML and will serve as a baseline for a more sophisticated system that will incorporate emotion recognition and generation. An experiment that will ask users for areas of improvement will be conducted. Special focus will be on shortcomings, if any, of the automated advisor's responses regarding trustworthiness and rapport.

### 1.2 Plans for Future Work

Several key factors are essential to the implementation of an persuasive spoken dialog system. In order to create such a system that is sensitive to user emotion, I will analyze different corpora and specifically focus on speech signals. I would like to obtain corpora that does not rely on actors. In (Craggs, Wood 2003), a corpus containing various emotional dialogs between nurses and a patient is described. In addition, the authors also provide guidelines for creating an annotation scheme that allows for ranking emotional levels. In (Forbes, Riley, Litman 2004), content in a corpus is labeled as having three basic types of emotions: positive, negative and neutral. However, if I were to adapt thier labeling method, modification to allow for finer-grained annotation would be necessary. I plan to detect and produce happiness, sadness, fear, surprise, and disgust. In addition, in the case of the graduate advisor, it is important to detect confusion.

The features of the speech signal that I will analyze for affective speech include prosody, loudness, energy, and tempo. Some machine learning techniques for emotion such as work conducted by Devillers, Vidrascu, and Lamel (2005) will be utilized and applied to my particular domain.

Recent work in persuasive textual dialog systems(Andrews, DeBoni, Manandar 2006) has shown interest in determining how to find a middle point between task-based dialog systems, which have very little emotion, and chatbots. Chatbots are interesting because, similar to the Eliza system, they solicit user emotion, but their weakness is the lack of modeling user state. I plan to extend his work and apply some of his methods to spoken dialog.

In order to achieve better trustworthiness and rapport with embodied conversational agents, the authors in (Cassell, Gill, Tepper 2007) built a model of deepening rapport based on data collected from studies between friends and strangers. Some of the verbal behaviors that were associated with friendship may be used to build a relationship in persuasive dialog.

## 2 Future of Spoken Dialog Research

In the future, I see users utilizing dialog systems in a variety of applications including health-care systems, video games, artificial entities such as robots, and command and control systems more naturally. Ideally, we will reach a point where third-party listeners will not know if someone is speaking to a human or a machine. This will be accomplished by allowing users to speak naturally, and have the dialog system understand not only the content, but also the implied meanings carried by acoustic features.

From the first introduction of Eliza the psychiatrist (Weizenbaum 1966) , which investigated human-computer interactions with dialog systems, it has been seen that computers can act as health-care agents. Creating systems that expand this simple echo response application into a more human-human conversation will benefit society.

Regarding virtual environments, both training and entertainment-based systems will be more effective. Soldiers can learn how to work with teammates quicker by working on building strong relationships by practicing with virtual agents. Future systems can be used to teach people how to gain rapport and how to better manage groups and become stronger leaders (persuasive dialog utilizing trust and rapport will help with this). Video games will become more realistic as the behavior of avatars will become closer to humans.

As pointed out by (Breazeal 2004), there will be a greater interest in robotic systems which are somewhat different from current human-machine interactions. Command and control of robotic systems and other machine technologies will benefit from better speech systems. There will also be a greater effort for multimodal interaction.

These advances can be accomplished by accurately modeling and reacting to the human mind. It is necessary to combine advances in machine learning, speech recognition and generation that incorporates emotion, cognitive state, physiological acts, and current work being done on textual analysis.

## 3 Suggestions for Discussion

- How to effectively detect and handle emotion, turn taking, back-channeling, response timing. What properties of speech can be extracted to predict these. How can this be done in real-time and what are some performance issues.

- Modeling the cognitive state based on acoustic features and content. How can dialog acts be recognized autonomously given a specific domain? Is personality a major factor in creating better dialog systems?

- How to use emotional speech to gain trust and rapport from a user and how this could effect the future of gaming systems and training.

- Collection of corpus dialogs with emotions from realistic dialogs as opposed to actors. How can we use information integration and the web to collect this data? What tools are best suited for analysis and labeling of different types of data?

## References

P. Andrews, M. De Boni, and S. Manandar. Persuasive argumentation in human computer dialogue. In *American Association for Artificial Intelligence*, 2006.

C. Breazeal. Social interactions in HRI: the robot view. *Systems, Man and Cybernetics, Part C, IEEE Transactions on*, 34(2):181–186, 2004.

Justine Cassell, Alastair J. Gill, and Paul A. Tepper. Coordination in conversation and rapport. In *Workshop on Embodied Language Processing*, pages 41–50, Prague, Czech Republic, June 28 2007. Association for Computational Linguistics.

R. Craggs and M. Wood. Annotating emotion in dialogue. SIGdial Workshop, 2003.

Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Netw.*, 18(4):407–422, 2005.

Kate Forbes-Riley and Diane Litman. Predicting emotion in spoken dialogue from multiple knowledge sources. In *HLT-NAACL 2004: Main Proceedings*, 2004.

J. Weizenbaum. ELIZAa computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.

## Biographical Sketch

Jaime Acosta is a PhD student in the Computer Science Department at the University of Texas at El Paso under the supervision of Dr. Nigel Ward. He has been employed at White Sands Missile Range, which is currently fully funding his studies, since 2004 and has since completed his Master's degree in Computer Science. His interests outside of school and work are video games, film, and playing guitar.

**Hua Ai**

University of Pittsburgh
5420 Sennott Square
210 S. Bouquet St.
Pittsburgh, PA 15260, USA

hua@cs.pitt.edu
www.cs.pitt.edu/~hua

## 1 Research Interests

My research interests lie generally in **spoken dialogue systems**, with a particular interest on **user simulation** for dialogue manager design.

### 1.1 Past, Current and Future Work

My research so far has focused on evaluating the utility of user simulation. We divided the evaluation problems into two parts: one is to estimate how humanlike the simulated corpora are; and the other is to measure how useful the user simulations are for a particular spoken dialog system design task since we believe this is a task-dependent question.

For the first part of the problem, we are interested in exploring automatic evaluation measures as well as verifying the validity of these automatic measures by means of a human assessment study. Previous research (Schatzmann et al., 2005) has proposed a group of automatic evaluation measures to distinguish between simulated corpora generated by different simulation models. We conducted a study (Ai and Litman, 2006) to examine the differentiating power of these evaluation measures to see to what extent they can distinguish between simulated corpora, between real corpora, and between simulated and real corpora. Our experiments show that some of these previously used measures do not provide enough information to figure out why two corpora are different; neither can they help us to draw conclusion on whether a corpus is a real corpus or simulated corpus. We observe that two real corpora can be very different when measured by these evaluation measures. Thus, even if these measures demonstrate that a simulated corpus is different from a real corpus, we cannot conclude that the simulated corpus is not realistic enough. We also conducted a human assessment study to validate these automatic measures using human judgments (Ai and Litman, 2008). We observe that it is hard for the human judges to reach good agreement when asked to rate the quality of the dialogs from given perspectives. However, the human ratings give consistent ranking of the quality of simulated corpora generated by different simulation models. We build a prediction model of human rankings using the automatic measures.

We notice that in other research fields (e.g. machine translation, document summarization) where automatic measures are used for evaluation, human assessment studies are also performed to validate the automatic measures. Furthermore, researchers are interested in finding correlations between the automatic measures and human assessment. We think this kind of validation and correlation study is also important for user simulation evaluation, but there is not much work done in previous research. We are currently carrying out a human assessment study to validate the previously proposed automatic measures.

For the second part of the evaluation problem, we are interested in studying the utility of different simulation models in the context of a particular dialog system design task. The two types of tasks we are currently focusing on are dialog strategy learning and dialog system evaluation. Reinforcement Learning (RL) is widely used to learn dialog strategy automatically (e.g., Singh et al., 1999, Henderson et al., 2005). Since RL training requires a large amount of training data, user simulation is considered as a promising approach to generate the large training corpus in a low-cost and time-efficient manner (Levin et al., 2000, Scheffler, 2002). However, it is unclear how realistic versus how exploratory a training corpus should be. In (Ai et al., 2007), we investigate what kind of user simulation is good for using Markov decision Processes to learn dialog strategies. In the study, we compare three simulation models which differ in their efforts on modeling the dialog behaviors in a training corpus versus exploring a potentially larger dialog space. Our results suggest that with sparse training data, a model that aims to randomly explore more dialog state spaces with certain constraints actually performs at the same or better than a more complex model that simulates realistic user behaviors in a statistical way.

For the task of dialog system evaluation, we hypothesize that a more realistic simulated corpus is preferable. Since the system strategies are evaluated and adapted based on the analysis of these simulated dialog behaviors, we would expect that these behaviors are what we are going to see in the test phase when the systems interact with human users. In (Ai and Litman, 2007), we propose a novel model to simulate student knowledge consistency in tutoring dialogs. This model constrains student performance on similar problems that requires similar knowledge based on the students previous performance while taking into account the learning effect of tutoring. We show that this new model does a better job in simulating the learning events happening in the tutoring sessions than a simpler model which generate user utterances in a probabilistic way. In our future work, we will further investigate whether this new simulation model is more helpful than the simple probabilistic model in system evaluation tasks.

## 2 Future of Spoken Dialog Research

A big challenge for the current spoken dialog systems is to handle speech recognition errors. On one hand, speech recognition errors are inevitable with the state-of-the-art speech recognizer. On the other hand, speech recognition rate are strongly correlate with user satisfaction. Therefore, how to detect speech recognition errors promptly and how to recover the errors in a natural way would be the very important issues to address in spoken dialog research.

## 3 Suggestions for discussion

Three possible topics for discussion could be:

Use and uselessness of user simulations: Best practices for building and evaluation user simulations? Possible criteria for a "good" user simulation?

Emotion detection in spoken dialogue systems: What kind of emotions should be detected in real applications? Is it possible to build up general emotional database to facilitate system development?

Multimodality: What are the strong points and short points for text and speech modalities? How to combine them in an efficient way? Should a system prompt the user to switch modality explicitly in some error conditions?

## References

H. Ai, and D. J. Litman. 2006. *Comparing Real-Real, Simulated-Simulated, and Simulated-Real Spoken Dialogue Corpora*. In Proc. of the AAAI Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems.

H. Ai, J. R. Tetreault, and D. J. Litman. 2007. *Comparing User Simulation Models for Dialog Strategy Learning*. In Proc. NAACL-HLT.

H. Ai, and D. J. Litman. 2007. *Knowledge Consistent* User Simulations for Dialog Systems. In Proc. Interspeech 2007.

H. Ai, and D. J. Litman. 2008. Assessing Dialog System User Simulation Evaluation Measures Using Human Judges. In Proc. of ACL 2008.

E. Levin, R. Pieraccini, and W. Eckert. 2000. *A Stochastic Model of Human-Machine Interaction For learning Dialogue Strategies*. IEEE Trans. On Speech and Audio Processing, 8(1):11-23.

K. Scheffler. 2002. *Automatic Design of Spoken Dialogue Systems*. Ph.D. diss., Cambridge University.

S. Singh, M. Kearns, D. Litman, and M. Walker. 1999. *Reinforcement learning for spoken dialogue systems*. In Proc. NIPS'99.

Henderson, O. Lemon, and K. Georgila. 2005. *Hybrid reinforcement/supervised learning for dialogue policies from communicator data.* In IJCAI Wkshp. On K&R in Practical Dialogue Systems.

J. Schatzmann, K. Georgila, and S. Young. 2005. *Quantitative Evaluation of User Simulation Techniques for Spoken Dialogue Systems*. In Proc .of 6th SIGdial Workshop on Discourse and Dialogue, 45-54.

## Biographical Sketch

Hua Ai is currently a 4th year Ph. D. student in the Intelligent Systems Program, University of Pittsburgh, USA. She works under the supervision of Dr. Diane J. Litman. She was born in P. R. China and received her B.S. in Computer Science from Shanghai Jiao Tong University, China. Her extra curricular interests include Traveling, Science Fictions, Dancing and Cartoons.

**Dan Bohus**

Microsoft Research
One Microsoft Way
Redmond, WA, 98052

dbohus@microsoft.com
http://research.microsoft.com/~dbohus

## 1 Research Interests

My current research agenda is focused on **situated natural language interaction**. The overarching question is: how can we develop systems that embed interaction and computation deeply into the natural flow of everyday tasks, activities and collaborations? More specifically, some of the areas and problems I am currently investigating are: **conversational scene analysis**, **multimodal sensor fusion, intention recognition, situated dialog and behavioral models, engagement models, self-supervised and lifelong learning and adaptation.**

### 1.1 Previous work

My previous work focused on **dialog management** and **error handling in task-oriented spoken dialog systems**.

As a platform for my dissertation research, I have developed RavenClaw (Bohus and Rudnicky, 2003), an open-source, reusable **dialog management** framework for task-oriented domains. RavenClaw has since been used to develop several systems spanning different domains and interaction types: information access (Bohus 2007, Raux et. al, 2006), browsing and guidance through procedural tasks (Bohus and Rudnicky, 2002), human-robot-interaction (Harris et al., 2006). Together with these systems, RavenClaw provides a robust basis for a number of research projects addressing issues such as error handling, multi-participant dialog, timing and turn-taking, dynamic generation of dialog plans, learning at the task level.

My dissertation work focused on the problem of **error handling** in task-oriented spoken dialog systems, and addressed three questions in this space: (1) how can a system reliably detect potential errors, (2) what strategies can be used to recover from different types of errors, and (3) how should a system choose between multiple such strategies at runtime.

With respect to **error detection**, the main contributions of my dissertation work were: an implicitly-supervised approach for training semantic confidence annotators (Bohus and Rudnicky, 2007); a data-driven approach for assessing the costs of various errors committed by a confidence annotator (Bohus and Rudnicky, 2005a); and a scalable belief updating framework for task-oriented spoken dialog systems (Bohus and Rudnicky, 2006). With respect to **error** recovery strategies, I have empirically investigated various types of non-understanding errors and several corresponding recovery strategies (Bohus and Rudnicky, 2005b). Finally, with respect to error recovery policies, I proposed and evaluated a supervised, online learning method for learning non-understanding recovery policies over a large set of recovery strategies (Bohus et al., 2006).

### 1.2 Current interests

Currently, my research agenda is focused on the notion of situated interaction. The central question is: how can we develop systems that embed interaction and computation deeply into the natural flow of everyday tasks, activities and collaborations?

To make this notion of situated interaction more concrete, consider the following examples. Several stores are currently experimenting with so-called “interactive billboards” that attract the attention of potential customers by detecting and responding to motion in their immediate vicinity (i.e. changing or animating their displays). It’s not too hard to imagine the day when such systems will become truly interactive; they will be able to grab your attention off the side-walk and engage you in a conversation about a product they’re trying to advertise and offer you coupons based on a drink they’re noticing in your hand. In the future, we can envision interactive systems that continuously monitor, assist and coordinate teams of experts through complex procedures and tasks (e.g. surgery, rescue operations, air traffic control, manufacturing etc.) Closer to home, they will perhaps watch over your shoulder and offer you guidance in the kitchen of the future as you experiment with a new recipe. The robots are well on their way! Today, they’re limited to very simple house-hold chores like vacuuming floors. Tomorrow, they will perform more complex tasks. We will encounter them monitoring patients in hospitals, giving directions in airports or stadiums, or perhaps serving as shopping assistants in the neighborhood mall. As the component technologies evolve, they will take on even more sophisticated roles such as educators, care-takers, and perhaps even social companions.

Bringing such systems into reality poses a number of interesting scientific and technological challenges. A common aspect in all of the scenarios above is that the interaction is **situated**, that is, it takes place in the real, physical world and is **deeply embedded in the natural flow of other human activities**.

Such systems will therefore need to be **situationally-aware**: they will have to fuse information from multiple sensors and knowledge sources, and continuously monitor and run inferences about what is happening in their surrounding environment. They need to be able to interact using **natural language**, and generate both verbal and non-verbal communicative behaviors that are in tune with social and cultural norms. Interaction planning will have to be tightly integrated with other complex subsystems for sensing, decision making and problem solving. Such systems should be able to engage in **mixed-initiative interaction** with one or **multiple participants**, and exhibit collaborative intelligence. Ideally, they should be able to **learn from their own experiences**, **adapt continuously throughout their lifetimes**, and **share the knowledge** they gain with each other.

## 2 Future of Spoken Dialog Research

In the past, a lot of applied dialog research was focused on telephone-based task-oriented spoken dialog systems. Today, a sufficient level of maturity has been reached in the field to allow for successful commercial development and deployment of these systems into daily use.

I believe that in the next 5-10 years more attention will shift towards issues in multi-modal, embodied and situated interactive systems, of the type described in the previous section. The challenges that lie ahead of us are exciting and many: situation awareness, multimodal sensor fusion, scene analysis, behavior and intention recognition, situated dialog management, situated grounding, engagement models, mixed-initiative and multi-participant interaction, life-long learning and adaptation.

## 3 Suggestions for discussion

- **challenges and opportunities in situated interaction.** Identify and discuss challenges in creating situated interactive systems: conversational scene analysis, intention recognition, behavioral models, situated dialog management, situated grounding, etc.
- **dialog management for human-robot interaction.** Discuss applicability and limitations of current dialogue management technologies (e.g. information-state, plan-based, POMDP-based, etc) in the context of human-robot interaction (e.g. multiple sensors, asynchronous events, multiple participants, etc.)
- **challenge problem(s) and evaluation**. Propose and discuss one or more challenge problem(s) for the field, and a corresponding evaluation process.

## References

Bohus, D.. 2007. *Error Awareness and Recovery in Conversational Spoken Language Interfaces,* Ph.D. Thesis, CS-07-124, Carnegie Mellon University, Pittsburgh, PA.

Bohus, D., Langner, B., Raux, A., Black, A., Eskenazi, M., Rudnicky, A., 2006. *Online Supervised Learning of Non-understanding Recovery Policies,* in Proceedings of SLT-2006, Palm Beach, Aruba.

Bohus, D., and Rudnicky, A. 2006. *A K-hypotheses + Other Belief Updating Model,* AAAI Workshop on Statistical Methods in Spoken Dialogue Systems, 2006.

Bohus, D., and Rudnicky, A. 2005b. *Sorry I didn't Catch That: An Investigation of Non-understanding Errors and Recovery Strategies,* SIGdial-2005, Lisbon, Portugal.

Bohus, D., and Rudnicky, A. 2005a. *A Principled Approach for Rejection Threshold Optimization in Spoken Dialog Systems,* Interspeech-2005, Lisbon, Portugal.

Bohus, D., and Rudnicky, A. 2003. *RavenClaw: Dialog Management Using Hierarchical Task Decomposition and an Expectation Agenda,* Proc. of Eurospeech'2003, Geneva, Switzerland; www.ravenclaw-olympus.org

Harris, T., Banerjee, S., and Rudnicky, A., 2005 *Heterogeneous Multi-Robot Dialogues for Search Tasks*, AAAI Spring Symposium: Dialogical Robots. Palo Alto, CA.

Raux, A., Bohus, D., Langner, B., Black, A., Eskenazi, M., 2006 – *Doing Research in a Deployed Spoken Dialog System: One Year of Let's Go! Public Experience*, Interspeech-2006, Pittsburgh, PA.

## Biographical Sketch

Dan is currently a researcher in the Adaptive Systems and Interaction group at Microsoft Research. His current research agenda is focused on situated interactive systems. Prior to joining Microsoft, Dan obtained his Ph.D. degree from Carnegie Mellon University, where he investigated problems of dialog management and error handling in task-oriented spoken dialog systems.

# Gwen Christian

Department of Linguistics
University of California Santa Cruz
1156 High Street
Santa Cruz, CA 95064-1077

jchristi@ucsc.edu

## 1 Research Interests

My research interests lie mainly in the optimization of **Human-Machine Interaction (HMI)**. I am interested specifically in the area of integrating expectation-management techniques into the designs of systems, and work centered on creating more conversational interfaces. Furthermore, I am also interested in the ability of automated spoken dialog systems to make possible new methods of communication between humans.

### 1.1 Expectation Management

The design of a spoken dialogue system necessitates managing the user's expectations to facilitate use. In Eldlund, Heldner & Gustafson (2006), the authors investigated the different 'metaphors' by which humans and machines interact. Using limited testing, they ascertained that one of the most important caveats of system design is an internally- consistent metaphor; that is to say that for a system using a human-like metaphor, human-like behavior is internally consistent; for what they call the 'interface' metaphor, machine-like behavior is internally consistent.

This was of particular relevance to me during my work as an intern on the EDAS (Emotive Driver Advisory System) with Ford Motor Company. Over the course of my work, I developed end-to-end implementation of dialogue functionality for two areas of in-car functionality: vehicle performance and entertainment. In the context of this, I was also responsible for modeling appropriate user-input and system-output, working within the human-like metaphor presented by the system as a whole. Over the course of my time there we were able to do limited testing during which it became clear that a more effective model could be generated. While I employed basic constraints on the system's lexicon and syntax to present dialogue to the user in terms that the system was able to understand, the testing made it clear that there were implicit assumptions at work that caused frequent frustration to our testers.

In future, it would be extremely worthwhile to do much broader testing in these terms in the context of a specific system, in order to create a more coherent interface that aligns itself more clearly with the expectations of the user. We can do this by considering not only what can be generated versus what can be recognized, but also non-verbal cues (for visual interfaces), indirect speech acts (c.f. Searle 1969), and overall system architecture.

### 1.2 Conversational Interfaces

In line with my interest in expectation management, I have great interest in the means by which more 'conversational' interfaces can be developed. This was also influenced by my work with Ford Motor Company. It would seem that one of the more important constraints in system design is allowing for the amount of attention the user is expected to expend over the course of an interaction, and to a greater extent, the sort of behavior they are willing to put up with and how to handle inevitable errors. (Martinovsky and Traum 2003).

For example, in call-routing or telephone systems it can reasonably expect that users are willing to devote their full (if limited) attention to the task at hand. However, when it is expected that the system will be used by drivers (to return to the EDAS example) an entirely different set of constraints become relevant. For drivers to navigate a lengthy menu-driven structure is not only unreasonable; it's unsafe. Therefore, there is an extent to which a more conversational interface is better, because it allows even naive users to operate in a mode in which they can be expected to be familiar with. However, consider the behavior of a helpful passenger; the context of being in the car functions as a constraint on the passenger's behavior, optimally not expecting the driver to respond to queries as quickly as might otherwise be expected. For these reasons, open mic (versus push-to-talk) technology is extremely useful, because it helps the system to be responsible for the timing of the conversation. Furthermore, studies could modify turn-taking algorithms to account for the vehicular context.

### 1.3 Spoken Dialog Translation

A final interest is the ways that spoken dialog systems can be used in terms of translation. I feel that this is one of the most useful areas of research in terms of people who could benefit from it. For example, the medical translation program, MedSLT[1] (Bouillion et al, 2007) is a spoken language translation system designed to facilitate a diagnosis between a patient and a doctor with no common language. A program proposed by a collaborator at University of Michigan is a translation system for use between Deaf and speaking people. Her proposal includes using existing technology like IBM's Say it Sign it (Tomasco 2007) (which takes speech input and produces signed output) and the MSignS project[2] (which works on visual recognition of ASL signs, functioning with about 90% accuracy with a vocabulary of roughly 100 signs (Clayton 2006)) and building an interface which translates both ways between signing and speaking participants. This is of great interest to me because it has the potential to change the way that humans communicate with one another.

## 2 Future of Spoken Dialog Research

Right now, I think that the average user considers spoken dialog systems in terms of push-to-talk user-operated systems, speaker-dependent dictation software, or menu-driven call-routing. As more and more interest seems to be exerted in the direction of creating more natural interactions, emotion-detection, and new systems which facilitate human-to-human communication, it seems that this will soon change. To return to the idea of the metaphor of a system, I speculate that we are shifting from a primarily 'interface' oriented perception, and into the realm of more 'human'- like behavior, which I perceive as having a lot of potential, but also many possible pitfalls. I think that this will present many ideas about relevant system metaphors by task, and also influence further research in both human-machine interaction, and human-human interaction.

## 3 Suggestions for Discussion

- Considering the metaphor of a system, what sorts of standardization could allow users to rapidly distinguish between 'interface' and 'human' systems? Could this stereotyping have negative effects on HMI?

- How can more effective turn-taking algorithms be constructed? What kinds of studies could provide the necessary insight to create contextual models? And would context-based turn-taking models be particularly effective in a broad sense, or would it make more sense to be as application-specific as possible?

- Machine translation: Translation is often considered to be more of an art than an algorithmic sorting pattern. What are the likely limits of machine translation and their future in general contexts or specific ones?

## 4 References

P. Bouillon, N. Chatzichrisafis, S. Halimi, B A Hockey, H. Isahara, K. Kanzaki, Y. Nakao, B. Novellas Vall, M. Rayner, M. Santaholma, and M. Starlander (2007). MedSLT: A Multi-Lingual Grammar-Based Medical Speech Translator. In *Proceedings of First International Workshop on Intercultural Collaboration*, IWIC2007, January 25-26, Kyoto, Japan.

B. Clayton (2006). MSigns: 3D Virtual Sign-Language Translation *Michigan Engineer* Spring 2006.

J. Edlund, M. Heldner, and J. Gustafson (2006). Two faces of spoken dialogue systems. In *Interspeech 2006 - ICSLP Satellite Workshop Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems*. Pittsburgh PA.

J. R. Searle. (1969) *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press, Cambridge, United Kingdom.

S. Tomasco. *IBM Research Demonstrates Innovative 'Speech to Sign Language' Translation System.* IBM Press Release. 12 Sep 2007 `http://www-03.ibm.com/press/us/en/pressrelease/22316.wss`

## Biographical Sketch

Gwen Christian is in the final year of her undergraduate studies at the University of California Santa Cruz, where she pursues degrees in Linguistics and Pre-and-Early Modern Literature. Her current research interests center around Human-Machine Interaction, influenced strongly from her time as a summer intern with Ford Motor Company, working on an in-car spoken dialog system. Her extracurricular interests include ballet, translation, and recreational reading.

[1] `http://www.issco.unige.ch/projects/medslt/index.shtml`

[2] `http://judyyu.net/msigns/index.html`

**Marcus Colman**

Interaction, Media and Communication
Department of Computer Science
Queen Mary, University of London
E1 4NS U.K.

marcus@dcs.qmul.ac.uk

## 1 Research Interests

My main research is in the area of **human-human interaction**; specifically in the way that **mutual-understanding** is achieved. I have examined this issue through the **quantification of mutual-engagement**. This was approached by developing coding protocols for the **Conversation Analysis** concept of **repair** and the broader concept of **ellipsis**.

### 1.1 A note on spoken dialogue research

As described above, my interests are in human to human communication rather than computer based systems. However, my research started as an approach to evaluating the efficacy of computer-mediated communication. The underlying process that needs to be examined is that one of understanding; it is for this reason that my research has a crossover with other developments in computer based dialogue systems.

### 1.2 Repair

'Repair' deals with problems in 'hearing' and understanding in communication. Previously this has only been looked at in a qualitative manner; through my work I have designed a reliable and valid coding scheme for examining this phenomenon.

Being able to fix a misunderstanding in communication is critical for all participants; statistically I have shown that there are two main points of interest. Firstly, there is a tendency for people to correct the problems with something that they themselves have said, and secondly that the likelihood of requesting a repair or correction will depend upon the role of the communicator - whether they are trying to explain something or understand something. When a request for a repair is made, it will generally appear as a choice of options rather than checking one specific understanding. Eye contact can affect the use of certain repair types; so can familiarity with the other participant. Healey et al. (2005) outlines the approach taken, although further unpublished work extends upon and supercedes this.

### 1.3 Ellipsis

As with the principle of repair, there was a question of whether mutual-understanding could be quantified and illustrated through the use of ellipsis (the omission of words or phrases from previous utterances). A protocol for identifying four types of ellipsis was developed and applied to a corpus. The four types (each divided into four subtypes) were anaphora; answers; questions and statements. Previous work has attempted to identify elliptical types of communication such as non-sentential utterances (e.g. Fernandez and Ginzburg; 2002), but without any inter-coder reliability - necessary for claims of coding scheme validity. What our research found was that significant differences exist between types of ellipsis, available to any participant in communication, are only used in certain cases. These differences have been found through coding a dialogue corpus, and various comparisons and contrasts to repair emerged. Simply put, any answers, questions or statements are statistically more likely to be elliptical if you have no eye contact with the other participant. The ellipsis coding scheme will be presented at this SIGdial conference; further publications will be forthcoming.

### 1.4 Applications of this research

Any approach to spoken dialogue systems must ultimately rely on how human communication works; it is this aspect that merges my work into computer based dialogue systems. The taxonomy of non-sentential utterances provided by Fernandez and Ginzburg (2002) demonstrates this to some degree; here I demonstrate a reliable, more comprehensive scheme that both can operationalise repair and ellipsis in communication.

## 2 Future of Spoken Dialogue Research

Current and future dialogue research: is an unknown quantity. If the question of intersubjectivity can be solved, then anything is possible.

## 3 Suggestions for discussion

- Can a computer follow anaphoric references if people have trouble following these?
- Is it necessary to be specific in requesting 'repairs' or further information? Or are 'wh-' words sufficient to explain the problem?
- How much information would be needed before an elliptical response can be constructed?

## References

Fernandez, R. and Ginzburg, J. 2002. Non-sentential utterances: a corpus study. *Traitement automatique des languages: dialogues* 43(2), 13-42.

Healey, P. G. T., Colman, M. and Thirlwell, M. 2005. Analysing Multi-Modal Communication: Repair-based measures of human communicative coordination. In *Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems* Dordrecht: Kluwer Academic.

## Biographical Sketch

Marcus Colman gained his BSc (Hons) in Psychology from Brunel University, and his MSc in Research Methods in Psychology from University College London. Currently he is carrying out research for his PhD at Queen Mary, University of London. He suffers from Asperger's syndrome and spends his free time chewing on his fingers and wishing he was somewhere else.

**María L. Flecha-García**

Department of Computer Science
University of Texas at El Paso
500 W University Avenue
El Paso TX 79968, USA

marisaflecha@gmail.com
www.cs.utep.edu/marisa

## 1 Research Interests

I am interested generally in **multimodal communication systems**, specifically those that use **embodied conversational agents** to communicate with the user. In relation to this, my main interest is the study of **natural dialogue** in **human-human interaction.** I am interested in the linguistic aspects of communication, such as **speech production and perception**, and **dialogue structure**, as well as in **non-verbal communication**, such as **body movement behaviour**. Another interest is the **collection and annotation of corpora** for studies of verbal and non-verbal behaviour that may help in the development of communication systems.

### 1.1 Past Work

During my M.Sc. studies at The University of Edinburgh, I became interested in the integration of different modalities in speech perception, and in particular in a well-known phenomenon by which, when presented with discrepant auditory and visual information, what you see affects what you hear. At the same time I was interested in foreign accent and face perception and so I ran a study to investigate whether ethnicity information perceived from a speaker's face could affect what the perceiver heard (Flecha-García 1998). I designed a perception experiment in which subjects were presented with audiovisual stimuli as follows. I created a synthetic speech continuum between the words "pip" and "peep" and I combined each of the items in the continuum with a digital image of one of three faces (as "speakers" of those words): one with typical male Anglo-Saxon features, one with male Mediterranean features, and a cartoon face as a control. The goal was to test whether subjects looking at a Mediterranean versus an Anglo-Saxon face would be biased to hear "peep" instead of "pip" due to the tendency for natives of Mediterranean countries, such as Spain and Italy, to pronounce both vowel sounds as "ee" in English. The results did not show the expected effect. One of the possible reasons was the use of unnatural stimuli, i.e. synthetic speech and still images. I became very aware of the importance of studying natural spontaneous speech and speech processing in real contexts.

With that in mind and keeping an interest in information from the speaker's face, for my Ph.D. thesis I studied the use of eyebrow raises in natural dialogue (Flecha-García 2006, 2007). The goal was to obtain information on when and why speakers raise their eyebrows during dialogues, which could be useful in the design of embodied conversational agents, for instance. In previous research eyebrow movements had mainly been studied in relation to the expression of emotion or as social signals. There were some indications that these movements could be related to the linguistic signal as well but these were mostly anecdotal observations, with very few exceptions, or were based on studies of synthetic speech. Thus, I did an empirical study of natural conversations by collecting and annotating a corpus of task-oriented dialogues. I tested several hypotheses about the relation of eyebrow raises to linguistic phenomena such as discourse structure, utterance function, information structure, and pitch accents. This revealed that the start of eyebrow raises was aligned with the start of accented syllables (i.e. where a pitch accent occurred). Also, eyebrow raises occurred more frequently at the beginning of certain high-level discourse segments than in other parts of the dialogue, and were longer and more frequent when giving instructions than in other types of utterances. Interestingly, speakers did not raise their eyebrows more often when asking a question nor when providing new (as opposed to given) information. I concluded that eyebrow raises are communicative, even if this might not be intentional, and that they may have two functions: structuring the spoken message and emphasizing parts of it. The results of this research could be useful not only for the design of embodied conversational agents, but also for dialogue systems capable of using non-verbal information from the users to interpret their intentions.

Between 1999 and 2005, I did some work as a Research Associate in different European projects studying spoken language. My main role was dialogue act annotation. I also trained and supervised a group of transcribers and annotators and adapted a tool for a project designed to determine whether feedback and time pressure affect what a speaker says (see e.g., Nicholson et al. 2003). As a Research Associate in a project at Oregon Health & Science University, I worked with a team in the prototyping of multimodal systems that use modalities such as speech, handwriting, and gestures in human-computer interaction in a meeting environment.

### 1.2 Current Work

I am now working in a team project on cultural differences in the way people behave in dialogue, including how and when the listener provides backchannel feedback showing attention, how speaker and listener communicate non-verbally, and how all this is interpreted by a listener or observer. We have already recorded an audiovisual corpus of conversations among and between Arabs and Americans and are currently annotating and analysing the data.

This work has direct applications to the field of spoken dialogue systems. For instance, if we can determine how cultures differ in the way they provide backchannel feedback, we will be able to implement this in the design of spoken systems so that they can respond appropriately to users of different cultures.

## 2 Future of Spoken Dialog Research

With no doubt, spoken dialogue systems will be more natural and efficient year by year and generations of new researchers will contribute to this development by bringing new ideas and perspectives. However, it is difficult to predict where this field will be in 10 years, as new and unexpected needs arise imposing new demands in the field of communication technology and its users. On the other hand, aside from the unexpected challenges of a changing world, something stays relatively unchanged and that is the basic way in which humans talk to other humans. We still have much to learn to fully understand the intricacies of this process, and that is why it is important that young researchers pay even more attention to the study of natural human communication behaviour, both verbal and non-verbal. In particular, an important issue to address is the development of standard, reliable methods and tools for the empirical study of human communication behaviour, especially multimodal behaviour.

## 3 Suggestions for discussion

- Tools and methods for the collection, annotation and analysis of natural multimodal communication behaviour: What is available? What do we need?
- The use of embodied conversational agents in spoken dialogue systems: what is needed for the design of agents that are helpful and not distracting
- Disfluency in users' speech: detection and handling.

## References

Flecha-García, María L. 1998. Investigation of a possible McGurk effect based on perceived ethnicity. Master's Thesis, The University of Edinburgh, September.

Flecha-García, María L. 2006. Eyebrow raising discourse structure and utterance function in face-to-face dialogue. In *Proceedings of* CogSci-2006, pages 1311-1316, Vancouver.

Flecha-García, María L. 2007. Non-verbal communication in dialogue: Alignment between eyebrow raises and pitch accents in English. In *Proceedings of* CogSci-2007, page 1753, Austin.

Nicholson, Hannele, E. G. Bard, A. H. Anderson, M. L. Flecha-García, D. Kenicer, L. Smallwood, J. Mullin, R. Lickley, and Y. Chen 2003. Disfluency under feedback and time pressure. *Proceedings of* EUROSPEECH-03, pages 205-208, Geneva.

## Biographical Sketch

María Flecha-García (Marisa) is a Post-doctoral Researcher in the Department of Computer Science in the **University of Texas at El Paso** working with Dr. David Novick and Dr. Nigel Ward. In 1998 she received a Master of Science degree in Cognitive Science and Natural Language from **The University of Edinburgh** (U.K.). She completed her doctoral degree in Linguistics at the same university in 2006 under the supervision of Dr. Ellen Bard and Professor D. Robert Ladd. Between 1999 and 2005 Marisa worked as Research Associate on various projects in the Department of Linguistics and what is currently called the School of Informatics at **The University of Edinburgh**. She was a Research Associate in the Department of Computer Science and Electrical Engineering at **Oregon Health & Science University** (2004-2005).

**Sudeep Gandhe**

Institute for Creative Technologies
University of Southern California
13274 Fiji Way, Suite #411d
Marina Del Rey, CA 90292

gandhe@ict.usc.edu
www-scf.usc.edu/~gandhe

## 1 Research Interests

I am interested in developing dialogue models with minimal annotation work. I have been exploring **statistical methods** that can bootstrap dialogue models from **un-annotated human-human dialogue** corpus. This work is focused on developing **Virtual Humans** with applications to simulation training, games, etc. I am also interested in developing technologies for Natural Language Dialog systems that would enable a **coherent interaction** between the user and the system.

### 1.1 Unsupervised Dialogue Modeling

Most dialogue systems operate on a dialogue-act level. That requires extensive annotation of dialogue corpora to learn how to convert from surface text to dialogue-acts. Moreover in models like the ones employing information-state updates, experts have to write down rules that operate on the dialogue act input and keep the information state updated. This semantic annotation and rule-writing is a bottleneck in rapidly developing dialogue systems for different domains. I am exploring statistical models that avoid this bottleneck by operating at surface text level (Gandhe and Traum, 2007). These methods are inspired from information retrieval and use un-annotated human-human dialogue corpus which is collected as part of role-plays and Wizard of Oz experiments. The basic assumption is that a dialogue can be carried out by retrieving the appropriate utterance from the corpus rather than constructing one from higher abstraction. We have found that such models are particularly useful for creating Virtual Humans for simulation training (Traum et. al., 2005) where the domain of interaction is controlled and the goal of the agent is to be as human-like as possible. I am also interested in evaluation of dialogue coherence models with minimal human input (Gandhe and Traum, 2008).

### 1.2 Coherent interactions

The idea of a question answering system, where answers are pre-recorded video segments has proved very useful in various applications for training and entertainment. Users are allowed to input a free text question which in turn elicits a pre-recorded video response. Although the video response tends to have very good value in terms of immersive experience, the very design of the system allows for a lack of coherence. Especially when there are no video responses directly answering the question or are not phrased in desired manner, the coherence gap problem is strikingly visible. We tried to address this issue by introducing short linking dialog between question and answer and thus bridging the gap. We carried out experiments to assess whether such linking dialogs can increase the coherence of interaction and proved that interactions with human-generated linking dialogs are statistically significant when compared to interactions without linking dialogs. (Gandhe et. al., 2004) Further analysis of human-generated linking dialogs reveals that these carry more information than present in the answer or the question. This leads us to realize the need for a knowledge base behind such a system. We have built such a knowledge base and have experimented with simple computer generated linking dialogs. (Gandhe et. al., 2006) have shown that these methods improve the perceived coherence of the interaction.

### 1.3 Speech to speech Translation

We developed a speech to speech translation system for medical domain. (Narayanan et. al., 2004) Using this system an English speaking doctor can communicate with a Farsi speaking patient and carry out the medical diagnosis. My work focused on GUI and the dialog manager. Only one participant, the doctor, can control the interaction. After speaking the utterance, the doctor is presented with multiple interpretations of that utterance and the doctor can choose one from those. The GUI also shows the history of the current dialog along with possible next utterances the doctor may choose to speak. The dialog manager component in this system is different from most of the dialog systems, in the sense that it has no active participation in carrying out the dialog. It can only assist the communication process. With this goal in mind, we split the dialog in phases. viz, introduction, registration, Q&A,

physical examination, diagnosis, conclusion. We also analyzed different medical cases, cardio, neuro, ent, ortho, ... Based dialog history and current phase and case estimations, the next possible doctor utterances are predicted and presented to the doctor for selection.

## 2 Future of Spoken Dialog Research

Currently speech as a modality is competing with more traditional interfaces like push buttons, GUI or touch screens. I believe in the next decade or so, speech interface in automobiles will be a commonplace occurrence. Speech technology has matured enough to provide a reliable alternative in cases where traditional interfaces are less desirable (e.g. while driving).

Looking more into the future, as spoken dialogue systems get more widespread, rapidly developing dialogue systems for new domains with minimum effort remains would an important challenge.

In non-task oriented systems building conversational partners or embedding these capabilities in other systems like car navigation can be an interesting goal to work towards for our times.

## 3 Suggestions for discussion

Possible topics for discussions:

- How to minimize the efforts required to rapidly develop dialogue systems for new domains? i.e. How to minimize the efforts of semantic annotation, dialogue act labeling?
- Evaluation: One of the recurring topics in previous discussions has been standardizing the evaluation of dialogue systems. Can an approach like shared task or competition for building dialogue systems be used to that effect? Can we get industry involved in providing funding and resources like data for such activities.

## References

David Traum, William Swartout, Jonathan Gratch, Stacy Marsella, Patrick Kenny, Eduard Hovy, Shri Narayanan, Ed Fast, Bilyana Martinovski, Rahul Baghat, Susan Robinson, Andrew Marshall, Dagen Wang, Sudeep Gandhe and Anton Leuski. 2005. *Dealing with Doctors: A Virtual Human for Non-team Interaction*. Demo to be presented at Sigdial 05.

S. Narayanan, S. Ananthakrishnan, R. Belvin, E. Ettelaie, S. Gandhe, S. Ganjavi, P. G. Georgiou, C. M. Hein, S. Kadambe, K. Knight, D. Marcu, H. E. Neely, N. Srinivasamurthy, D. Traum, and D. Wang. 2004. *The Transonics Spoken Dialogue Translator: An aid for English-Persian Doctor-Patient interviews.* in working notes of the AAAI Fall symposium on Dialogue Systems for Health Communication, pp 97--103.

Sudeep Gandhe, Andrew Gordon, Anton Leuski, David R Traum, and Douglas W. Oard. 2004. *First steps toward linking Dialogues: mediating between free-text questions and pre-recorded video answer*. Presented at the Army Science Conference.

Sudeep Gandhe, Andrew S. Gordon, David Traum. 2006. *Improving question-answering with linking dialogues.* International Conference on Intelligent User Interfaces (Sydney, Australia, Jan 29 - Feb 1, 2006).

Sudeep Gandhe and David Traum. 2007. *Creating Spoken Dialogue Characters from Corpora without Annotations.* Proceedings of Interspeech 2007.

Sudeep Gandhe and David Traum. 2008. *Evaluation Understudy for Dialogue Coherence models.* Proceedings of SIGdial 2008.

## Biographical Sketch

He is currently pursuing a PhD degree in Computer Science at University of Southern California, Los Angeles. He works at Institute for Creative Technologies under the advisement of Dr. David Traum. Before that he has earned Masters in Computer Sceince from USC and B. Eng. (Computer) from V.J.T.I, Mumbai University, India. His interest lie in developing techologies for unsupervised dialogue modelling using unannotted corpus.

Milica Gašić

Department of Engineering
University of Cambridge
Trumpington street
Cambridge CB1 1PZ
mg436@cam.ac.uk

## 1 Research Interests

My research interest lies in the area of **statistical** approaches to **Dialogue Management**. I am primarily focused on **Partially Observable Markov Decision Process** framework to Spoken Dialogue Systems. This framework has the potential to enable learning from data, learning from interaction with both simulated and real users and, also, to be robust to recognition errors.

### 1.1 Modelling Dialogue as a Partially Observable Markov Decision Process

One of the main problems that real world applications of Spoken Dialogue Systems are encountering is the difficulty to deal with the errors from the speech recogniser in a noisy environment. As almost any real application is not noise free, this is a large obstacle preventing dialogue systems from wide use.

It has been suggested that the Partially Observable Markov Decision Process (POMDP) can provide a principled mathematical framework to deal with the uncertainty that originates from errors in the speech recognition by retaining and updating the distribution over all states in each turn (SJ Young, 2002; JD Williams and SJ Young, 2007a). The idea is that the dialogue state is hidden and therefore the policy should not be based on a single state but on the distribution over all possible states.

The main weakness of POMDP approach is its computational intractability for even very simple domains. However, it has been suggested that factoring the states into summary states can enable using POMDPs for policy optimisation (JD Williams and SJ Young, 2007a; JD Williams and SJ Young, 2007b). It has also been shown that the policy learning can be performed online in interaction with a simulated user (Thomson et al., 2007).

### 1.2 Current Work

My current work involves training a POMDP-based dialogue manager with a simulated user in a noisy environment and investigating how that influence the performance with real users in noise.

My work is focused on the **Hidden Information State** system (HIS). What makes HIS different to other POMDP-based dialogue managers is its representation of user goal in the dialogue state. It retains the full representation of user goal though partitions of the user goal space. The partitions are created based on the information that the user has provided. After each dialogue turn the partitions and their probabilities are updated and pruning is performed to the ones with a low probability (SJ Young et al., 2007; Thomson et al., 2007; JD Williams and SJ Young, 2007b).

In order to reduce the dimensionality, the state space (the master space) is mapped into a much smaller summary space. Reinforcement learning is performed on the probability distribution over the summary space - the belief space. Since the belief space is continuous, it is dicretised into grid points. Then, an online batch policy iteration is performed on these points. Finally, the outcome of the policy is mapped back into the master space.

The results of my current work suggest that the training in noise leads to better performance at higher semantic error rates then in training in noise free environment. Also, the ability of a dialogue manager to make use of N-best inputs from a recogniser improves in the performance at higher error rates.

### 1.3 Future Work

What I encountered during the training of the dialogue manager with a simulated user is training a very small number of grid points can reach the same performance as using a larger number of grid points for learning. This suggests that there the current choice of summary space can be enriched in order to be more informative.

On the other hand, a rich summary space can lead to computational intractability due to the exponential nature of the algorithms used in the POMDP framework.

Therefore, my future work will consist of investigating different summary spaces and trying to develop a technique for defining an optimal summary space that best represents the whole state space on one hand and that is enables computationally tractable training on the other hand.

On a longer run, I would like to be able to investigate more advanced policy optimisation techniques, as well as the potential for online learning with real users.

## 2 Future of Spoken Dialog Research

In the next decade, I expect noise robustness issues that prevent dialogue systems from a wider application to be resolved. Also, it is very important to define the framework which is domain independent and easily transferable to other domains. POMDPs seem to have the potential for this, but additional research is needed in order to to further investigate the possibilities to obtain desirable results with a computationally tractable approximations.

It would be desirable that the process of building a Statistical Spoken Dialogue System in the next decade becomes fully automatic. This would mean that the simulated user can be trainable from real data, that the dialogue manager can learn from the interaction with the simulated user and finally that is also able to learn with real users. Therefore the future of the Dialogue Manager is highly dependent on the simulated user.

## 3 Suggestions for Discussion

- Can Spoken Dialogue Systems be robust to ASR input errors?
- How predictive is the dialogue performance of a dialogue manager trained on a simulated user for its performance with real users?
- Which policy optimisation algorithms could be applicable to dialogue manager learning with a simulated user?
- What issues need to be solved to enable online learning with real users?

## References

B Thomson, J Schatzmann, K Weilhammer, H Ye, and SJ Young Training a real-world POMDP-based Dialog System In *HLT/NAACL Workshop ''Bridging the gap: Academic and Industrial Research in Dialogue Technologies''*, Rochester

JD Williams and SJ Young Partially Observable Markov Decision Processes for Spoken Dialog Systems *Computer Speech and Languages*, 21(2):393-422

JD Williams and SJ Young Scaling POMDPs for Spoken Dialog Management *IEEE Audio, Speech and Language Processing*, 15(7):2116-212

SJ Young, J Schatzmann, K Weilhammer, and H Ye. The Hidden Information State Approach to Dialog Management In *ICASSP 2007*, Honolulu, Hawaii

SJ Young Talking to Machines (Statistically Speaking) In *Int Conf Spoken Language Processing*, Denver, Colorado.

## Biographical Sketch

After graduating in Computer Science and Mathematics from the University of Belgrade in 2006, Milica Gašić did MPhil in Computer Speech, Text and Internet Technology at the University of Cambridge. She completed the course in 2007 with the thesis *Limited Domain Synthesis of Expressive Speech*. In October 2007 she became a PhD candidate in Dialogue Modelling under the supervision of Prof Steve Young. She is working in the Dialogue Systems Group, `http://mi.eng.cam.ac.uk/research/dialogue/`.

# Alexander Gruenstein

M.I.T. Computer Science and
Artificial Intelligence Lab
32 Vassar Street
Cambridge, MA 02139, USA

alexgru@csail.mit.edu
http://www.mit.edu/~alexgru/

## 1 Research Interests

The primary motivation underlying my research is to create multimodal interfaces which make complex tasks easier and more natural. I aim to build flexible, fluid, and intelligent interfaces which leverage speech. Humans use language to effortlessly share complex ideas, and research in spoken dialogue systems provides a critical foundation for advancing both human-computer interaction and artificial intelligence. To that end, I am particularly interested in making access to dialogue systems ubiquitous: I endeavor to advance dialogue systems from their current role as the annoying voice on the other end of the phone line to that of a nearby companion, with a rich multimodal interface.

My general methodology is to first develop prototype multimodal interfaces, which I then use to perform research into ways in which we can improve users' experiences with such systems. It is critical, however, that we get these prototypes in the hands of users, so that we can understand how to improve their experiences with the systems we develop. To that end, I have been developing a web-based platform with which compelling and interactive multimodal dialogue systems can be developed, deployed, and iteratively improved. This makes it possible to build compelling multimodal interfaces using web standards, allowing us bring multimodal interfaces out of the lab and into the hands of users on their desktop, laptop, and tablet computers—and, increasingly, on their mobile devices.

Much of my current work is centered around a map-based multimodal interface I've developed called *City Browser*, which provides restaurant, museum, and public transportation information for several metropolitan areas in the United States (Gruenstein et al., 2006; Gruenstein and Seneff, 2006). In developing *City Browser*, we have worked towards both *portability* and *scalability*: the framework is portable to any map-based system, and it is scalable in that it can easily accommodate large numbers of restaurants in any number of metropolitan areas. Recently, I have also been involved in developing a multimodal home entertainment system, which allows users to access their media content on their television, via their mobile device (Gruenstein et al., 2008). In the past, I have contributed to a multimodal interface for commanding a robotic helicopter (Lemon and Gruenstein, 2004; Lemon et al., 2002).

I am particularly interested in the ways in which conversational context can be used to improve the performance of multimodal interfaces. A major usability problem that such systems face is that users (especially those unfamiliar with speech recognition) may not be sure what they can say, and—at the same time—when they do speak within the confines of the system's understanding capability, they may often be misunderstood due to speech recognition errors. Much of my recent research has been focused around mitigating these difficulties via context-sensitive techniques. In particular, I have looked at ways to use contextual information to dynamically update **context-sensitive language models** (Gruenstein and Seneff, 2006; Gruenstein et al., 2005; Lemon and Gruenstein, 2004), contextually **shape user utterances** (Gruenstein and Seneff, 2007), and **assign confidence scores** to system responses (Gruenstein, 2008).

Finally, I also have an interest in exploring **incremental speech processing** in multimodal applications. In particular, I'm interested in ways that the graphical modality can be used to give feedback to a user as he or she speaks, so that the user understands immediately what the system has heard so far, and how it has updated its beliefs as a result. I have only done some initial exploration in this area, but hope to soon study it in more detail.

## 2 Future of Spoken Dialog Research

I believe that, in the near future at least, mobile devices are the ideal target for spoken dialogue systems. Mobile devices are becoming increasingly powerful, and now offer a wide range of functionality previously available only on desktop computers. However, they still have extremely limited input capabilities: small screens and tiny keyboards. There has been an enormous increase in interest in providing speech interfaces to mobile devices, however these tend usually to focus on form-filling types of approaches where users may use speech to fill in one field of a form at time. While this is a good starting point, smarter interfaces which allow for more complex interaction should be the ultimate goal.

Mobile devices are also interesting because while speech may be a very good input modality, it is not always the ideal output modality. Mobile devices are used in many contexts: while driving, walking, running, riding the train, etc. As such, the ideal output modality may shift depending on what the user is doing. While driving, speech output may be quite useful; but while sitting on a train, the screen might be the best way to give information. Moreover, users keep mobile devices with them, so they may simply want a response to a query stored for later reference, rather than described to them immediately. A mobile device which has some awareness of the context in which it is being used, and of the user's personal preferences and habits, can enable very interesting, long-ranging interactions.

## 3 Suggestions for Discussion

**Multimodality and Incrementality** It can be challenging, both technically and theoretically, to develop multimodal interfaces which incrementally understand a user's utterance as he or she speaks. How quickly, and in what manner, could a multimodal interface respond graphically to a user during an utterance. Can this be done in a way that is non-obtrusive, yet useful?

**Life partners** We tend to think of dialogue systems as applications that a user will interact with for a brief period of time, before moving back to "real life." What about a dialogue system (or systems) which play a meaningful role of people's lives day in and day out? What is involved in making speech a part of people's daily interactions with computers? What does a dialogue system mean in this context? How can a speech interface be more like a "life partner" which helps you accomplish all sorts of tasks throughout your day? There are all sorts of aspects to this problem, from hardware to client/server architectures, to AI and HCI. Moreover, it highlights a real need to perform longitudinal studies of how people make use of speech interfaces when they can use them frequently.

**Generic dialogue systems** One of the holy grails of dialogue systems research has always been to make truly "reusable" components, which application designers can integrate into their applications without any knowledge of speech or dialogue systems technology. We have VXML and SALT, but where should we head next? What capabilities and knowledge could researchers in our field provide as APIs to people who know very little about speech or dialogue systems technology? Integrating speech technology into an application is not as simple as letting a user speak whenever they see a text box (or is it?). What would be more useful? What would be interesting enough to get application developers to notice?

## References

Alexander Gruenstein and Stephanie Seneff. 2006. Context-sensitive language modeling for large sets of proper nouns in multimodal dialogue systems. In *Proc. of IEEE/ACL 2006 Workshop on Spoken Language Technology*.

Alexander Gruenstein and Stephanie Seneff. 2007. Releasing a multimodal dialogue system into the wild: User support mechanisms. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 111–119.

Alexander Gruenstein, Chao Wang, and Stephanie Seneff. 2005. Context-sensitive statistical language modeling. In *Proc. of INTERSPEECH*, pages 17–20.

Alexander Gruenstein, Stephanie Seneff, and Chao Wang. 2006. Scalable and portable web-based multimodal dialogue interaction with geographical databases. In *Proc. of INTERSPEECH*.

Alexander Gruenstein, Bo-June (Paul) Hsu, James Glass, Stephanie Seneff, Lee Hetherington, Scott Cyphers, Ibrahim Badr, Chao Wang, and Sean Liu. 2008. A multimodal home entertainment interface via a mobile device. In *Proc. of the ACL Workshop on Mobile Language Processing*.

Alexander Gruenstein. 2008. Response-based confidence annotation for spoken dialogue systems. In *Proc. of the 9th SIGdial Workshop on Discourse and Dialogue*.

Oliver Lemon and Alexander Gruenstein. 2004. Multithreaded context for robust conversational interfaces: context-sensitive speech recognition and interpretation of corrective fragments. *ACM Transactions on Computer-Human Interaction*, 11(3):241–267.

Oliver Lemon, Alexander Gruenstein, and Stanley Peters. 2002. Collaborative activities and multi-tasking in dialogue systems. *Traitment automatique des langues*, 43(2):131–154. Special issue on dialogue.

## Biographical Sketch

Alexander Gruenstein is a Computer Science Ph.D candidate at MIT, where he is in the Computer Science and Artificial Intelligence Laboratory. He received B.S. and M.S. degrees in Symbolic Systems from Stanford. Before coming to MIT, he worked at *BeVocal* on commercially available dialogue systems systems, as well as at Stanford's Center for the Study of Language and Information. Last summer he worked at *vlingo* on speech interfaces for mobile devices, and this summer will work with the speech group at *Google*.

# Stefan Hamerich

Harman/Becker Automotive Systems GmbH
CoC Speech & Connectivity
Speech Services
89077 Ulm – Germany

shamerich@harmanbecker.com
www.hamerich.de/stefan

## 1 Research Interests

My research interests lie generally in the area of **spoken dialogue systems**, with a special focus on **automotive** systems. I am especially interested on **dialogue descriptions** and **speech output** of these systems.

After getting in contact with the wide field of speech and language in university I was a working student at IBM European Speech Research creating a phone-based dialogue system informing on traffic jams presented in (Günther et al., 2000). Since I was really fascinated being able to make machines speak with humans and (at least mostly) understand their answers, I got stuck within the field of speech dialogue systems. Therefore I built a speech-based shopping system during my diploma thesis (Hamerich, 2000). Within this work I learnt a lot about dialogue strategies and which impact they have on the dialogue flow and user's behaviour. And I made my first own system evaluation, which was sometimes really frustrating, since real users behaved very different from what I expected.

After finishing my studies I worked within the EC funded project GEMINI. The aim of the project was to develop a platform able to generate a dialogue flow for accessing data from a database. For making the platform work a special abstract dialogue description language was created (Hamerich et al., 2003) which even could be used independently of the GEMINI project (Schubert and Hamerich, 2005). The platform developed within the project was able to cover multi-modal error handling (Wang et al., 2003). The final version even covered mixed-initiative dialogues in multiple languages and offered user-modelling and overanswering (Hamerich et al., 2004). Within the project I was mainly occupied with the definition of the abstract dialogue modelling language, with the general design of the generation platform and the specialties of error handling for the speech modality. Finally the project was successfully finished; a complete overview is given in (D'Haro et al., 2006).

At Harman/Becker I am working on automotive dialogue systems, which as embedded systems are in some aspects really different from telephone-based ones, refer to (Hamerich and Hanrieder, 2004; Hamerich, 2005) for further details.

As a member of the dialog research and tools team within Harman/Becker I was dealing with new prototypes of advanced applications for in-car use. One of these prototypes I was working on was a system able to do music selection in a car based on spoken artist or title names. A first approach has been presented in (Wang et al., 2005), a more extended and elaborate version has been presented in (Wang and Hamerich, 2008).

Currently I am more involved into product development, where my main task is to care about (synthesised) speech output for speech dialogue and navigation systems in cars.

Another task is my Ph.D. thesis, which I am doing in my spare time under the supervision of Walther von Hahn from the University of Hamburg and Wolfgang Minker from the University of Ulm. Within the scope of my thesis I am working – based on my work with GEMINI – on DiaGen, a tool to support development for automotive speech dialogue systems. The tool will be presented in (Hamerich, 2008). A first prototype created with DiaGen has already been presented in (Hamerich, 2007).

## 2 Future of Spoken Dialog Research

Regarding dialogue research in general I think technology already is quite good, but nevertheless most applications are not yet good enough to be used by inexperienced users. From my point of view the problem results in the design of most systems. Most novices using a SDS for the first time just do not know what to say. If the system detects the timeout it is important to just guide these users as good as possible. Here some systems just tend to repeat the initial question. This of course does not help at all. It seems to me that most research is still dealing with technological issues, which of course are needed and helpful. But at a certain step each technology is to be used in an application. And that is when soft facts like dialogue design (here especially prompt design) become important, because we need users to use our systems. Unfortunately the perfect design for each situation has not yet been found. Therefore dialogue design is still a thing to be discussed. Personally I think this is a field which needs much more research and attention by researchers.

I would wish that these general things are investigated with more efforts and we all get closer to more useable systems. We all shall not forget that finally someone has to buy our systems. And there are less people just buying new technology because it is cool (so called early adaptors) and there are more people only buying things if they have a sense. For SDS that sense could mean faster access to information, more comfortable device control, or just an intuitive user interface for everyone. And this goal still is far away. So let's go to reach that goal!

## 3 Suggestions for Discussion

I propose the following topics for discussion:

- System evaluation: which experiences do we have, which metrics do we use, which standards could be created?
- Dialogue design: how to write/develop a good and usable speech dialogue? What are best practices? Which experiences do we have?
- Image of speech: if talking to people not involved into speech, they do not like speech systems at all due to bad experiences or whatever. What can we do to change that and maybe have a bright future with speech?

## References

L.F. D'Haro, R. de Córdoba, J. Ferreiros, S.W. Hamerich, V. Schless, B. Kladis, V. Schubert, O. Kocsis, S. Igel, and J.M. Pardo. 2006. An Advanced Platform to Speed up the Design of Multilingual Dialog Applications for Multiple Modalities. *Speech Communication*, 48(8):863–887.

C. Günther, S.W. Hamerich S., Kunzmann, and T. Roß. 2000. ISA: A Traffic Jam Information System based on the IBM ViaVoice Telephony Toolkit. In *Proceedings of the Workshop 'Voice Operated Telecom Services'*, pages 63–66, Ghent, Belgium. COST 249.

S.W. Hamerich and G. Hanrieder. 2004. Modelling Generic Dialog Applications for Embedded Systems. In *Proc. ICSLP*, pages 237–240, Jeju, Korea.

S.W. Hamerich, Y.-F.H. Wang, V. Schubert, V. Schless, and S. Igel. 2003. XML-Based Dialogue Descriptions in the GEMINI Project. In *Proceedings of the Berliner XML-Tage*, pages 404–412, Berlin, Germany.

S.W. Hamerich, R. de Córdoba, V. Schless, L.F. d'Haro, B. Kladis, V. Schubert, O. Kocsis, S. Igel, and J.M. Pardo. 2004. The GEMINI Platform: Semi-Automatic Generation of Dialogue Applications. In *Proc. ICSLP*, pages 2629–2632, Jeju, Korea.

S.W. Hamerich. 2000. Strategien für Dialogsegmente in natürlichsprachlichen Anwendungen [in German]. Master's thesis, Computer Science Department, University of Hamburg, Hamburg, Germany.

S.W. Hamerich. 2005. Speech Dialogue Systems for Cars - an Overview. *SDV – Sprache und Datenverarbeitung*, 29(2):107–118.

S.W. Hamerich. 2007. Towards Advanced Speech Driven Navigation Systems for Cars. In *Proc. IE*, pages 247–250, Ulm, Germany.

S.W. Hamerich. 2008. From GEMINI to DiaGen: Improving Development of Speech Dialogues for Embedded Systems. In *Proc. SIGdial*, Columbus, USA.

V. Schubert and S.W. Hamerich. 2005. The Dialog Application Metalanguage GDialogXML. In *Proc. EUROSPEECH*, pages 789–792, Lisbon, Portugal.

Y.-F.H. Wang and S.W. Hamerich, 2008. *Dybkjær, L. and Minker, W. (Eds.): Recent Trends in Discourse and Dialogue*, chapter Designing Speech-Controlled Media File Selection for Automotive Systems, pages 25–43. Springer, Dordrecht, Netherlands.

Y.-F.H. Wang, S.W. Hamerich, V. Schubert, V. Schless, and S. Igel. 2003. Multi-Modal and Modality Specific Error Handling in the GEMINI Project. In *Proc. ISCA Workshop on Error Handling in SDS*, pages 139–144, Chateau d'Oex, Switzerland.

Y.-F.H. Wang, S.W. Hamerich, M.E. Hennecke, and V.M. Schubert. 2005. Speech-controlled Media File Selection on Embedded Systems. In *Proc. SIGdial*, pages 217–221, Lisbon, Portugal.

## Biographical Sketch

Stefan Hamerich is working with speech-controlled infotainment systems in cars. He sees a lot of nice, expensive cars with cool new technology and gets paid for this by Harman/Becker, which is the automotive division of Harman International.

He received a master's degree in computer science (speech and language oriented AI) from the University of Hamburg, Germany in 2001. In 1999 and 2000 he was a working student at the IBM European Speech Research group in Heidelberg, Germany. In his spare time he is working on his nearly finished Ph.D. thesis dealing with a tool to ease development of user-friendly speech dialog systems for automotive environments.

**Anna Hjalmarsson**

Speech, Music and Hearing
School of Computer Science and Communication
Lindstedtsvägen 24
100 44 Stockholm

annah@speech.kth.se
http://www.speech.kth.se/~annah

## 1 Research Interests

My research focuses on developing **flexible spoken utterance generation**. Humans produce speech incrementally and on-line as the dialogue progresses using information from several different sources in parallel. We anticipate what the other person is about to say in advance and start planning our next move while this person is still speaking. When starting to speak, we typically do not have a complete plan of how to say something or even what to say. Yet, we manage to rapidly integrate information from different sources in parallel and simultaneously plan and realize new dialogue contributions. Pauses, corrections and repetitions are used to stepwise refine, alter and revise our plans as we speak (Clark & Wasow, 1998). My work so far has revolved around three different dialogue systems:

### 1.1 DEAL

DEAL is a dialogue system for second language conversational training in Swedish, currently being developed at KTH (Hjalmarsson et al., 2007). DEAL is a multidisciplinary research platform for exploring challenges and potential benefits of combining elements from computer games, dialogue systems and language learning. From dialogue research point of view this approach contributes with several novel and interesting objectives and challenges. These include how to design dialogues which are fun and natural using a language which suits the vocabulary and language complexity of language learning students on various levels. Since efficiency and task completion are no longer the main objectives, dialogue systems in a serious game context do not have to be predictable, rational or even co-operative. Instead, we need to consider how to build systems which are fun, educational and addictive to talk to. I am involved in various parts of the DEAL project but I am mainly working on the implementation of human-like generation strategies.

In order to generate output in a more stepwise manner extended knowledge on how to signal relations between different segments of speech is needed. In a recent data collection effort human-human dialogue data was recorded and labelled to extend the knowledge of human interaction and in particular to distinguish different types of cue phrases used in the DEAL domain. Ten different functional cue phrases (see Hjalmarsson, in press for more details) were labelled with high agreement between labellers (kappa coefficient 0.82, p=0.05). The data is a valuable resource of information for the use of cue phrases in the DEAL domain as well as how they are lexically and prosodically realized.

### 1.2 The KTH Connector

During 2005-2007 I was involved in the development of the KTH Connector, a spoken dialogue system acting as a personal secretary. Within this project I conducted an experimental study (Hjalmarsson & Edlund, in press) in order to see how human-like variability is perceived in the context of a spoken dialogue system. Human-human dialogue data were collected and used to simulate a system with human-like variability by replacing one of the human parties in the recordings with a synthetic voice. The task was for non-participating listeners to compare two different versions of system behaviour in a dialogue. One version was an imitation of a human behaviour and one version was constrained to contain less variability. The results support that the system version based on a human speaker was perceived as more human-like, polite and intelligent compared to a system version with less varying behaviour.

### 1.3 AdApt

My master's thesis project was a data collection of human-machine dialogue which was used as a basis for a PARADISE evaluation and a qualitative study of various dialogue parameters in AdApt (Hjalmarsson, 2005). The aim of the AdApt project was to study human-computer interaction in a multi-modal conversational dialogue system. The tasks of the system are associated with finding available apartments in Stockholm.

## 2 Future of Spoken Dialog Research

The research issues of today, including how to detect and recover from speech recognition errors, perform user evaluations and develop components which can be used across applications, will likely be urgent for many years. However, if future dialogue systems are used within a wider range of applications, including systems for entertainment and tutoring, these issues needs to be reconsidered. Depending on which type of dialogue systems we aim at, we need to find out what is crucial for that particular type of system. We also need methods to evaluate single components, components in combination as well as entire systems at different stages of system development (Edlund et al., in press).

## 3 Suggestions for discussion

- How can we build dialogue systems which give the users a sense of talking to a system with human-like conversational capabilities? If aiming at spoken dialogue systems coherent with a human metaphor, i.e. which evoke us to talk to it as talking to another human, we might need to consider how to build dialogue systems that model an extended set of human dialogue manners including grounding strategies, incremental processing and disfluencies such as repetitions, hesitations and restarts.
- Human computation has shown to be an effective way to collect large amounts of data for various tasks for which humans are better suited than computers. Good examples of such experiments are the ESP game (http://www.espgame.org/, von Ahn) and Paek et al. (2007). Much of the effort when building spoken dialogue systems lies in colleting speech and human strategies within the target domain. What type of data can be collected and how can we make people provide data for free and because it is fun for our particular purposes?
- How can we test theories of interaction without building an entire dialogue system? Empirical experiments to study particular issues are not always studied within fully working dialogue systems. How can we best benefit and what are the pitfalls of methods like Wizard of Oz studies and analysis of human-human interaction?

## References

Clark, H. H., & Wasow, T. 1998. *Repeating words in spontaneous speech*. Cognitive Psychology, 37(3), 201-242.

Edlund, J., Gustafson, J., Heldner, M., & Hjalmarsson, A. in press. *Towards human-like spoken dialogue systems*. To be published in Speech Communication, Special Issue on Evaluating new methods and models for advanced speech-based interactive systems.

Hjalmarsson, A., & Edlund, J. in press. *Human-likeness in utterance generation: effects of variability*. To be published in Proceedings of the 4th IEEE Workshop on Perception and Interactive Technologies for Speech-Based Systems. Kloster Irsee, Germany.

Hjalmarsson, A., Wik, P., & Brusk, J. 2007. *Dealing with DEAL: a dialogue system for conversation training*. In Proceedings of SigDial (pp. 132-135). Antwerp, Belgium.

Hjalmarsson, A. 2005. *Towards user modelling in conversational dialogue systems: A qualitative study of the dynamics of dialogue parameters*. In Proceedings of Interspeech. Lisbon, Portugal.

Hjalmarsson, A. in press. *Speaking without knowing what to say... or when to end.* To be published in Proceedings of SIGDial. Columbus, Ohio, USA.

Paek, T., Ju, Y-C., & Meek, C. 2007. *People Watcher: a game for eliciting human-transcribed data for automated directory assistanc*e. In Proceedings of Interspeech (pp. 1322-1325). Antwerp, Belgium.

## Biographical Sketch

Anna Hjalmarsson received a M.A. in Cognitive Science from Linköping University in 2003. In 2004 she was accepted to the Graduate School of Language Technology, GSLT, and became a PhD student at Speech Music and Hearing, KTH, under the supervision of Professor Rolf Carlson. Anna's research focuses on developing flexible and context sensitive generation in spoken dialogue systems. Anna has also been working on the CHIL project an Integrated Project (IP 506909) under the European Commission's Sixth Framework. She is also working with the development with the language training game DEAL.

# Preethi Jyothi

The Ohio State University
395 Dreese Laboratories
2015 Neil Avenue
Columbus, Ohio 43210-1277

jyothi@cse.ohio-state.edu
www.cse.ohio-state.edu/~jyothi

## 1 Research Interests

My research interests broadly cover **machine learning in spoken dialogue systems** with a special focus on how an interactive system adapts to the environment it is deployed in and how it automatically learns from its own experiences without any explicit help from the developers of the system. I am particularly interested in learning more about techniques that help spoken language interfaces to perform in a robust manner and error-handling strategies that would improve the efficiency of the dialogue between the system and the user. I am in the first year of my doctoral studies and I am still exploring more about my intended area of research.

Even with the current state of the art Automatic Speech Recognition (ASR) systems, ASR errors are inevitable and they significantly compromise the end-to-end performance of the spoken dialogue system (SDS).In their interactions with SDSs, people try to correct these ASR errors in dialogue turns subsequent to the one where the ASR error occurred. (Litman et al., 2006) describe an analysis of the various features which distinguish the above mentioned user corrections from other user input and discusses the use of machine learning techniques to identify such user corrections.

Spoken Dialogue Systems usually have little generalization power and are not portable across application domains. (Ammicht, Fosler-Lussier and Potamianos, 2007) implement application-independent semantic and pragmatic modules which can be used in dialogue system design. This includes an introduction of a domain-independent semantic representation of the user input that efficiently represents semantic ambiguity.

ASR confidence scores serve as an initial estimate of the interpretation error of the user's input. Ideally, an SDS should consolidate information from subsequent user responses to improve and update the scores of the hypotheses assigned to a particular concept in the system. (Bohus and Rudnicky, 2006) propose the use of machine-learning to address this problem. The authors make use of a representation which keeps track of the k-best hypotheses for a given concept at a given time and train a linear regression model to construct the updated beliefs.

## 2 Future of Spoken Dialog Research

Speech is the most natural medium for human-human communication and is increasingly used these days to interact with computer agents in the form of spoken dialogue systems. This is also aided by the remarkable growth of the basic infrastructure for communications technology, even in the developing countries.

I believe the main challenge that faces dialogue systems today is the ability of the system to automatically optimize its responses while incorporating contextual information which is learnt from interactions with the user. This problem is further complicated by the limitations of current speech technologies, especially for recognition. Thus, in order to make the human-machine interaction successful, another hurdle that current dialogue systems need to cross is to devise effective strategies for determining when a dialogue is not proceeding as expected and choosing a mechanism to deal with the error once it is detected.

## 3 Suggestions for Discussion

- Current spoken dialogue systems are not sufficiently robust and one of the most important challenges facing these systems is their lack of flexibility when faced with understanding errors. What are the various error-detection and repair strategies that could be employed after the ASR stage to improve the end-to-end performance of these systems?

- The purpose of a practical dialogue is to achieve a concrete goal. The language used in practical dialogue, while large, is small relative to the full complexity of human discourse. (Allen, et al., 2000) People tend to stay within this subset of language when their main focus is on achieving a stated objective. Most of the current dialogue systems are largely goal-oriented and thus support practical dialogues. How can we use machine learning techniques (Reinforcement Learning, Markov Decision Processes) to help such systems enable dynamic contextual interpretation and generation of responses?

## References

Diane Litman, Julia Hirschberg and Mark Swerts. 2006. *Characterizing and Predicting Corrections in Spoken Dialogue Systems*. Computational Linguistics, 32(3):417–438.

Dan Bohus and Alex Rudnicky. 2006. *A "K Hypothesis + Other" Belief Updating Model*. AAAI Workshop on Statistical and Empirical Approaches to Spoken Dialogue Systems, Boston, MA.

Egbert Ammicht, Eric Fosler-Lussier and Alexandros Potamianos. 2007. *Information Seeking Spoken Dialogue Systems - Part 1: Semantics and Pragmatics*. IEEE Transactions on Multimedia, 9(3):532-549.

James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu and Amanda Stent. 2000. *An architecture for a generic dialogue shell*. Natural Language Engineering, 6(3-4):213-228.

## Biographical Sketch

Preethi Jyothi is a first year PhD student in the Department of Computer Science and Engineering at The Ohio State University. She is working under the advisement of Dr.Eric Fosler-Lussier in the Speech and Language Technologies Laboratory.

**Rohit Kumar**

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15217

rohitk @ cs.cmu.edu
www.rohitkumar.net

## 1 Research Interests

I am investigating the use and the effectiveness of rich behavioral strategies for conversation. In particular, I am interested in the use of **social conversational behavior in multi-user task domains** like collaborative learning. Some interesting characteristics of social behavior include: several interpersonal aspects including **emotion, personality and social role** comprise social behavior; social behavior like greetings can often be detached from task oriented behavior suggesting **reusability of behavior** across task domains; interplay of social and task oriented behavior presents interesting issues related to **prioritization and coordination** of conversational behavior.

### 1.1 Past work

In our work, we have developed tutorial dialogue systems for domains such as thermodynamics, calculus, middle school math, earth sciences, physics, psychology, civics, etc., but our most long term and focused effort has been in the context of the CycleTalk project (Rosé et al., 2006). In this project, we have evaluated the use of Conversational Agents playing the role of tutors in learning environments. These learning environments are interfaces like instant messengers and chat rooms and include one or more students working on a given exercise while the tutor helps the students.

In the process of development and evaluation of tutorial dialog systems, we have identified requirements for a framework suitable for building such systems. We have also identified behavioral shortcomings on the agent's part, having which may help in their pedagogical effectiveness.

**Framework Requirements** Conversational Agents used for educational applications need to be integrated with variety of learning environments ranging from chat rooms to simulators to virtual worlds. Agents built for such applications need to be able to detect and respond to events like problem completion, incorrect use of concept, etc. This calls for a need to have a framework with elegant and scalable integration of procedural as well as event driven behavior. Finally, such agents should have decomposable behavior to allow reusability and ease distributed development.

**Behavioral Shortcomings** While agents acting as tutors are effective at helping students learn (Kumar et al., 2007), we observe that the students often ignore the tutors in the presence of another human in setting like collaborative learning. We also observe that the students, who participate in more instructive conversational turns with the tutor, learn more. This led us to hypothesize that tutors capable of exhibiting social behavior to engage the students may be more effective.

### 1.2 Ongoing work

Recently, we have developed Basilica as a framework for developing conversational agents. The underlying motivation of this framework was to serve the requirements I have listed earlier. Basilica takes a decomposition stance on development of intelligence. Each Intelligent behavior that is part of a conversational agent is identified and developed as a separate component. We argue that this approach enables re-use of behavioral components. Also, this facilitates integration by isolating the components required for the integration of the agent with external environments.

**Basilica: A new framework** Basilica (Kumar et al., Submitted 2008) is an event-driven framework, which enables development of conversational agents by using two basic types of behavioral components, namely Actors and Filters. The components communicate using Events. The Actor component, as the name suggests displays behavior. Filters on the other hand observe behavior. Behavior and data are encapsulated into Events. We adopt a programmatic approach to development compared to other frameworks which use proprietary specification and scripting languages. Each component is a Java class. We preserve the benefits of using a scripting language by building wrapper components. We have built a wrapper component for scripts developed using the TuTalk authoring tools (Cui and Rosé, 2008). More details on the intricacies of this framework are beyond the scope of this paper.

**Social Behavior and its Effect** Based on the finding about the shortcomings of agent behavior, we used

Basilica to implement new social behavior alongside the instructional behavior already exhibited by the older versions of our conversational agents. We have found that social conversational behavior exhibited by agents is effective in changing the user behavior towards the agents (Chaudhuri et al., 2008) as well as their behavior towards each other (Kumar et al., 2007). However, this kind of behavior has not been successful at influencing the outcome measures related to task success. Currently, our hypothesis is that the social behavior in conversation causes a change in the interpersonal dynamics between the interlocutors. This change is measurable by questionnaires and conversation analysis methods. On the other hand, the influence of the change in interpersonal dynamics on learning takes time and an experiment conducted over longer duration of time may be required to find this effect.

## 2 Future of Conversational Agent Research

Conversational agent and dialog system research in the near future may move into the development of systems which interact with more than one user at a time. The role of the conversational agent may shift from being an active participant to an occasional contributor for some of these applications. The emergence of such agents may be driven by the popularity of multi-user virtual environments like Second Life. In such environments and applications, the agents may need to exhibit behavior other than typical task oriented behavior in order to maintain its presence among the users (Kumar et al., Submitted 2008).

The switch from single-user to multi-user scenarios as well as the need for social behavior alongside task oriented behavior will present new and interesting challenges to this research community. Some of these challenges include coordination and prioritization between different behavioral components, design and evaluation of the strategies used to create social behavior, identification and adoption of participant roles, tracking change in interpersonal dynamics, etc.

I believe one way to approach these issues would be to develop experimental applications in which multiple users would participate in a useful activity while being supported by one or more conversational agents.

## 3 Suggestions for discussion

- Emerging application domains where multiple users and agents can work together.
- The role of the agents in such domains.
- Frameworks support required for development of agents in such domains.

## References

Carolyn P. Rosé, Rohit Kumar, Vincent Aleven, Allen Robinson, Chih Wu. 2006. *CycleTalk: Data Driven design of Support for Simulation based Learning*, Intl. Journal on Artificial Intelligence in Education Special Issue on the Best of ITS 2004, 16, 195-223.

Rohit Kumar, Gahgene Gweon, Mahesh Joshi, Yue Cui, and Carolyn P. Rosé. 2007. *Supporting Students Working Together on Math with Social Dialogue*, SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA

Rohit Kumar, Carolyn P. Rosé, Mahesh Joshi, YiChia Wang, Yue Cui, and Allen Robinson. 2007. *Tutorial Dialogue as Adaptive Collaborative Learning Support*, Intl. Conf. on Artificial Intelligence in Education, Los Angeles

Rohit Kumar, Baba Kofi A. Weusijana, and Carolyn P. Rosé, 2008 (submitted), *Building conversational agents in Second Life using Basilica*, Interspeech 2008, Brisbane, Australia

Sourish Chaudhuri, Rohit Kumar, and Carolyn P. Rosé. 2008. *It's Not Easy Being Green: Supporting Collaborative Green Design Learning*, Intl. Conf. on Artificial Intelligence in Education, Montreal

Yue Cui and Carolyn P. Rosé. 2008. *An authoring tool that facilitates the rapid development of Dialogue Agents for Intelligent Tutoring Systems*, Intl. Conf. on Artificial Intelligence in Education, Montreal

## Biographical Sketch

Rohit Kumar is a Ph.D. student at the Language Technologies Institute at Carnegie Mellon University. He received his Masters in Language Technology from Carnegie Mellon in 2007 and Bachelors in Computer Science and Engineering from Punjab Engineering College, India in 2003. As a research scientist at International Institute of Information Technology, Hyderabad, India, he worked on developing state of the art text to speech systems for Indian languages. Since joining Carnegie Mellon University in 2005, Rohit has worked with Carolyn P. Rosé on the development and evaluation of tutorial dialog systems. He also worked on the ConQuest spoken dialog systems with Alexander I. Rudnicky and led the deployment of ConQuest at IJCAI 2007. As a student leader at Carnegie Mellon, Rohit is a Graduate Student representative for his department and serves on the finance committee. He also serves on the advisory committee for the Indian Graduate Students Association.

**François Mairesse**

Department of Engineering
University of Cambridge
Trumpington street
Cambridge CB2 1PZ
francois@mairesse.co.uk
mi.eng.cam.ac.uk/˜farm2

# 1 Research Interests

My research interests lie generally in models of individual differences in language, with a special focus on **user modelling**, **stylistic natural language generation** and theories from **personality psychology**.

## 1.1 Past work

I've recently completed my Ph.D. thesis on statistical models for detecting and conveying linguistic variation in dialogue systems. Although there are many ways to express any given content, most dialogue systems do not take linguistic variation into account in both the understanding and generation phases, i.e. the user's linguistic style is typically ignored, and the style conveyed by the system is chosen once for all interactions at development time.

Over the past few years, psychologists have identified the main dimensions of individual differences in human behaviour: the Big Five personality traits (Norman, 1963). The Big Five traits are hypothesised to provide a useful computational framework for modelling important aspects of linguistic variation. My thesis first explores the possibility of recognising the user's personality using data-driven models trained on essays and conversational data (Mairesse et al., 2007). I then tested whether it is possible to generate language varying consistently along each personality dimension in the information presentation domain. I implemented PERSONAGE: a language generator modelling findings from psychological studies to project various personality traits (Mairesse and Walker, 2007). PERSONAGE was used to compare various generation paradigms: (1) rule-based generation, (2) overgenerate and select and (3) generation using parameter estimation models—a novel approach that learns to produce recognisable variation along meaningful stylistic dimensions without the computational cost incurred by overgeneration techniques. These generation methods were evaluated based on human judgements, showing that human judges can detect the personality conveyed by the system's utterances, even if multiple traits are projected simultaneously (Mairesse and Walker, 2008).

## 1.2 Current and future work

While my previous work has focused on dialogue system components (user modelling and natural language generation), I recently joined Cambridge's Dialogue Systems group in order to develop statistical models for semantic parsing and language generation, and to test these components within a fully trainable dialogue system. This work is part of the CLASSiC project fudned by the European Union. In the future, I am hoping to test various hypotheses about personality-based alignment in different dialogue tasks, in order to evaluate the benefits of modelling user characteristics linguistic cues, as well as to assess what style or personality is the most suitable given a specific dialogue application (e.g. information presentation, intelligent tutoring, interactive drama).

# 2 Future of Spoken Dialog Research

I believe that modelling linguistic variation can greatly improve the interaction in dialogue systems, such as in intelligent tutoring systems, video games, or information retrieval systems, which all require specific linguistic styles. Previous work has shown that linguistic style affects many aspects of users' perceptions, even when the dialogue is task-oriented (Cassell and Bickmore, 2003; Wang et al., 2005). Moreover, users attribute a consistent personality to machines, even when exposed to a limited set of cues (Reeves and Nass, 1996), thus dialogue systems manifest personality whether designed into the system or not.

I thus believe that the future of dialogue system research is to produce systems that can control that variation in a scalable way. This scalability will require moving towards data-driven approaches at all levels of the output generation, i.e. controlling the level of initiative, the language generation process and the speech synthesis module in a consistent way. While previous research has produced highly trainable components (e.g. speech recogniser, text-to-speech engine), young researchers will need to focus on how to make the language understanding, dialogue management and language generation components re-usable across domains.

## 3 Suggestions for Discussion

- Dialogue system personalisation: is it important to model the system's output style, and if so what kind of style or personality should the system convey for different dialogue applications (e.g. intelligent tutoring systems, tourist information presentation, financial information retrieval)?

- In the near-future, will data-driven methods remove the need for rule-based knowledge in all dialogue system components?

- What issues need to be solved before all dialogue system components become truly re-usable from one domain to the other (e.g. dialogue manager, language generator)?

## References

J. Cassell and T. Bickmore. Negotiated collusion: Modeling social language and its relationship effects in intelligent agents. *User Modeling and User-Adapted Interaction*, 13:89–132, 2003.

F. Mairesse and M. A. Walker. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 496–503, 2007.

F. Mairesse and M. A. Walker. Trainable generation of Big-Five personality styles through data-driven parameter estimation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008.

F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research (JAIR)*, 30:457–500, 2007.

W. T. Norman. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality rating. *Journal of Abnormal and Social Psychology*, 66:574–583, 1963.

B. Reeves and C. Nass. *The Media Equation*. University of Chicago Press, 1996.

N. Wang, W. L. Johnson, R. E. Mayer, P. Rizzo, E. Shaw, and H. Collins. The politeness effect: Pedagogical agents and learning gains. *Frontiers in Artificial Intelligence and Applications*, 125:686–693, 2005.

## Biographical Sketch

François Mairesse is a research associate at the University of Cambridge, working in the Machine Intelligence Lab as part of the EU CLASSiC project. He recently completed his Ph.D. thesis at the University of Sheffield supervised by Prof. Marilyn Walker, focusing on models of individual differences in dialogue systems. The thesis investigates techniques for the recognition of the personality of the user as well as the control of the personality conveyed by the system. In 2006, he did an internship at AT&T Labs in Florham Park working on paraphrase acquisition from web reviews. In 2004, he obtained a Master's degree in Computer Science and Engineering from the Université Catholique de Louvain in Belgium.

# Matthew Marge

Carnegie Mellon University
Language Technologies Institute
School of Computer Science
Pittsburgh, PA 15213

mrmarge@cs.cmu.edu
www.cs.cmu.edu/~mrmarge

## 1 Research Interests

My research interests lie in the domain of **spoken dialogue systems** for **human-robot interaction**. As part of the TeamTalk project, I work on improving how multimodal robots can better interpret spatial language and their environment. Our project's domain requires that humans and robots work together on a "treasure hunting" task. I rely on the principle of **grounding** for my research efforts. Currently, I am exploring the capabilities of spatial perspective-taking in **human-robot dialogue**.

### 1.1 Spoken Dialogue Systems for Surveys

My undergraduate research included building a spoken dialogue system for course-based surveys with the assistance of colleagues in the Computer Science Department at Stony Brook University. My chief responsibilities were to design, implement, and evaluate an automated course evaluation system called the "Rate-A-Course System" (Stent et al., 2006). This system allowed people to set their own goals for the conversation, such as how many topics to discuss about a course and in what order to discuss them. For my honors thesis, I designed and conducted experiments with human participants to study how speech-driven dialogues can adapt to users. Our project goal was to understand what dialogue designs were most effective when interacting with users that have specific goals in conversation.

### 1.2 Adaptive Human-Robot Dialogue

In past research, I investigated the cognitive and social aspects of robotics at the Carnegie Mellon University Human-Computer Interaction Institute. I studied how robots might adapt to the existing knowledge of novices and experts via dialogue interaction. Pearl, an interactive robot from the CMU Robotics Institute, was used in our experiments. We investigated whether, given a topic of conversation such as cooking, Pearl should use technical terms with experts but longer explanations of those terms with novices. One of my responsibilities was to enhance Pearl's existing adaptive dialogue system. I did this by increasing the number of appropriate responses Pearl could give to participants' questions about cooking tools. When we compared responses of novices and experts, we found that novice cooks appreciated Pearl and performed best when it gave detailed explanations of the tools rather than the tool names alone (Torrey et al., 2006). By contrast, expert cooks found Pearl patronizing when it gave them detailed explanations of the tools.

As an intern at the Naval Research Laboratory, I developed a scenario for disambiguating dialogue in a human-robot team (Fransen et al., 2007). In this scenario, two people are directing a mobile robot to retrieve an item from a dangerous area. In order to improve the ability of the robot to disambiguate dialogue spoken by the team members, I worked with graduate students to integrate several functionalities into the robot, including gesture recognition, sound localization, and natural language understanding.

### 1.3 Spatial Human-Robot Dialogue

Currently, I am working with other members of the TeamTalk project and the larger Boeing TreasureHunt project under the supervision of Dr. Alex Rudnicky. This project investigates how robots can better collaborate with humans using speech and dialogue with the goal of finding "treasures" in a real-world location. I helped prepare the current version of the virtual TeamTalk system, along with a research associate and a summer intern. This simulation, built using the USARSim system, provides us a vehicle to test our spoken dialogue system without the need to manage actual robots (Balakirsky et al., 2006). The system is built using the RavenClaw/Olympus Dialogue Architecture (Bohus et al., 2007).

I am currently exploring the capabilities of spatial perspective-taking in human-robot dialogue. We want to find out how spatial perspective-taking, both in reference to members of a human-robot team and to objects in the environment, can be incorporated into the current TeamTalk platform. This included developing a small scenario that we could begin to study. I am also learning more about the dialogue concept of grounding for this work.

Our exploration of spatial perspective-taking in human-robot dialogue has led us to design an experiment to assess how humans give simple dialogue commands in reference to members of a human-robot team. We have developed this experiment in order to conduct it formally with human participants. I have also conducted a literature review of spatial reasoning, spatial language, and its applications in human-robot interaction.

### 1.4 Future Work with Human-Robot Dialogue

We intend to expand the experimental stimuli from snapshots of scenarios to real-time interaction with the TeamTalk virtual system in the near future. After I conduct a formal experiment on spatial perspective-taking in human-robot scenarios, I plan to develop the TeamTalk dialogue engine with a spatial reasoning component. I intend to base it on what humans typically say when giving spatial commands from my experiment. Once we are established on a spatial reasoning component that can refer to team members in a scenario, I plan to extend the component to reference objects in the environment of the human-robot dialogue scenario.

## 2 Future of Spoken Dialog Research

Interactive spoken dialogue applications will steadily increase in popularity over the next decade. This is dependent upon public acceptance of these systems, which have often frustrated the general public due to poor design and implementation. We should see a decrease in the number of call centers answering customer service phone lines, for example. I also expect that as the field of robotics expands, dialogue will become an increasingly greater need given the wide range of tasks that robots can perform. Our generation of spoken dialogue researchers should be able to develop formalized, accepted methods for evaluating spoken dialogue systems. Also, our generation should be able to analyze and improve the fine-grained aspects of human-computer dialogue to improve everyday interactions with public users. I think that grounding will become an even more ever-present concept that dialogue researchers must rely on when designing effective spoken dialogue systems. Attaining these goals requires that formal user studies be performed with publicly accessible spoken dialogue systems. In addition, these goals will require our generation of dialogue researchers to educate incoming students in our departments about the importance of spoken dialogue applications and the need for more researchers in the field.

## 3 Suggestions for Discussion

- Cognitive plausibility: What is the cognitive plausibility of existing spoken dialogue systems? Do they manage dialogue in a way that the cognitive science community would find acceptable?
- Environment interaction: How can we have spoken dialogue systems better process changing information about the environments of their applications?
- Spoken dialogue design: How can we best minimize the chance that speech recognition errors irritate the users of our systems? Will this require more careful design of dialogue prompts?
- Appealing to the next generation: What kind of applications should we discuss with new members of our departments to excite them about dialogue research? Should we be discussing new “killer apps”?

## References

S. Balakirsky, C. Scrapper, S. Carpin, and M. Lewis. 2006. Usarsim: providing a framework for multirobot performance evaluation. In *Performance Metrics for Intelligent Systems Workshop (PerMIS)*, pages 98–102.

Dan Bohus, Antoine Raux, Thomas K. Harris, Maxine Eskenazi, and Alex Rudnicky. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *HLT-NAACL 2007 Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technology*.

B. Fransen, V. Morariu, E. Martinson, S. Blisard, M. Marge, S. Thomas, A. Schultz, and D. Perzanowski. 2007. Using vision, acoustics, and natural language for disambiguation. In *2nd ACM/IEEE Conference on Human-Robot Interaction*, pages 73–80.

Amanda Stent, Svetlana Stenchikova, and Matthew Marge. 2006. Dialog systems for surveys: The rate-a-course system. *Spoken Language Technology Workshop, 2006. IEEE*, pages 210–213.

Cristen Torrey, Aaron Powers, Matthew Marge, Susan Fussell, and Sara Kiesler. 2006. Effects of adaptive robot dialogue on information exchange and social relations. In *1st ACM SIGCHI/SIGART Human-Robot Interaction Conference*, pages 126–133.

## Biographical Sketch

Matthew Marge is a PhD student in the Language Technologies Institute at Carnegie Mellon University. He received the MSc degree in Artificial Intelligence specializing in Natural Language Engineering from the University of Edinburgh in 2007. Prior to graduate school, he received the BSc degree in Computer Science and Applied Mathematics from Stony Brook University in 2006. He is currently funded by the Boeing Treasure Hunt Project and a National Science Foundation Graduate Research Fellowship. Matthew is a former recipient of a St. Andrew’s Society of the State of New York Scholarship and an Edinburgh-Stanford Link studentship. Some of his hobbies include squash, bowling, and travelling.

## Zeljko Medenica

University of New Hampshire
33 College Road
Durham, NH 03824
USA

zeljko.medenica@unh.edu
http://zeljko.medenica.googlepages.com/

## 1 Research Interests

My research is related to **spoken human-computer interaction (SHCI) in vehicles**, with particular emphasis on the influence of different aspects of the SHCI and other factors on **driving safety** and **performance**.

### 1.1 Motivation

The motivation for this research comes from two prominent characteristics of our society: driving and using in-car devices. Given the large amount of time that people on average spend behind the wheel (Lowe, 2005) and the increasing availability of computing resources that can be used inside a vehicle, many companies have introduced a number of mobile devices for drivers into the consumer market. Some wide known examples are cell-phones with hands-free dialing, GPS navigation devices, live traffic reports, in-car multimedia, etc.

### 1.2 Past, Present and Proposed Future Work

Since we are exposed to this trend of increasing number of in-vehicle devices, the question that naturally comes to mind is how drivers cope with this? With more and more states adopting legislation barring the use of mobile devices in cars (Governors Highway Safety Association, 2008), speech user interfaces that allow drivers to dial numbers or control multimedia hands-free are becoming increasingly common. One example is Ford Sync (Ford Motor Company, 2008) that allows drivers to make phone calls and control their music selection hands-free. Also, in our work at the University of New Hampshire, we have deployed an in-vehicle speech user interface for the Project54 system (Kun et al., 2004). The Project54 system integrates multiple devices in a police cruiser such that they can be controlled using voice commands. Speech was always anecdotally associated with the safest way of communicating with in-vehicle devices. In order to confirm this, we compared the influence of the manual user interface and the Project54 speech user interface while interacting with the police radio on driving performance (Medenica and Kun, 2007). Our results showed that manual interaction decreased driving performance significantly, while speech interaction did not. But, in our recent study (Kun et al., 2007), we have explored the effects of speech recognizer accuracy, press-to-talk (PTT) button and dialogue repair of an in-car speech user interface on driving performance. Our results showed that poor speech recognition performance negatively influences driving performance. We also found that when the speech recognition performance is poor, having to use a PTT button results in worse driving performance than when using ambient recognition. This stands as a proof that there is no absolutely "safe" technology and that further studies are necessary to investigate the possible negative effects that those may introduce.

The main problem here is that the effect of different in-car technologies on the psychology and driving performance of users has not been adequately addressed in the research literature. A related problem is that of determining how to integrate these technologies so as to reduce the threat of accidents. Ideally, speech interaction should not introduce any impairment to the primary visual and cognitive task of driving. However, as was shown in our study above, this is not always the case. Our research endeavors to address such questions using a state-of-the-art driving simulator, an eye-gaze tracker, and physiological metrics.

One important question is how should in-car technologies be assessed for their impact on driving performance? In the past, different researchers used different variables, such as reaction time (Chisholm et al., 2007), crash and near-crash rates (Neale et al., 2005), etc. However, because of different experimental conditions and used measurements, it is very difficult to compare and reuse different studies and to draw clear conclusions from them.

Our goal is to create standards for evaluating in-car user interfaces and to use these standards to evaluate the influence of the characteristics of speech user interfaces on driving performance. In order to make the results easy to interpret and compare, we will design a set of standardized obstacle tests in a driving simulator that ordinary drivers would be able to navigate without accidents, if they focus their attention on the driving task. This way we would be able to precisely measure the impact of in-car devices and their interfaces on driving performance. Our hypothesis is that there are three factors (and their interactions) that will have an effect on driver's response in the obstacle test:

- Level of visual and cognitive load caused by different road conditions,
- Speech user interface characteristics (recognition accuracy, interface type – command-and-control or natural language processing, PTT button location),
- Psychological state - driver emotions combined with speech interaction (Nass and Brave, 2005).

Understanding these characteristics will not only be of scientific interest, but will also be of practical value to vehicle user interface designers. Thorough analysis of these characteristics will reveal which of them are statistically significant, which would enable us to develop statistical models that would predict situations when breakdowns in driving performance are likely to occur.

## 2 Future of Spoken Dialog Research

We are witnessing an almost exponential increase of the number of vehicles on our roads. Also, technological advances are introducing more and more electronic devices into vehicles that are supposed to make time spent in vehicles more enjoyable and productive. Looking at this trend we can expect cars to become "docking stations" where users would be able to add their own devices as they desire. Speech has proven to be a very efficient way of handling these technologies, but more research has to be done in this area to eliminate potential harmful interference of the characteristics of the speech user interface and other factors that can impair the main task of driving.

## 3 Suggestions for discussion

Some of the topics for discussion are:

- Ways of designing speech-user interfaces in vehicles,
- How to incorporate affective cues into speech-user interface system,
- Influence of emotions on subject's perception of the speech-user interface quality.

## References

Andrew L. Kun, W. Thomas Miller, III, William H. Lenharth, 2004. *Computers in Police Cruisers*, IEEE Pervasive Computing

Andrew L. Kun, Tim Paek, Zeljko Medenica, 2007. *The Effect of Speech Interface Accuracy on Driving Performance*, Interspeech, Antwerp, Belgium

Clifford Nass, Scott Brave, 2005. *Wired for Speech*, Cambridge, MA: MIT Press

Ford Motor Company, 2008. http://syncmyride.com

Governors Highway Safety Association, 2008. *Cell Phone Driving Laws*, www.ghsa.org/html/stateinfo/laws/cellphone_laws.html

Shelly Lowe, 2005. *Many Workers Have Long Commutes to Work*, US Census Bureau Press Release

Susan L. Chisholm, Jeff K. Caird, Julie Lockhart, Lisa Fern, Elise Teteris, 2007. *Driving Performance while Engaged in MP-3 Player Interaction: Effects of Practice and Task Difficulty on PRT and Eye Movements*, Driving Assessment, Stevenson, WA

Vicki L. Neale, Thomas A. Dingus, Sheila G. Klauer, Jeremy Sudweeks, 2005. *An Overview of the 100-Car Naturalistic Study and Findings*, National Highway Traffic Safety Administration

Zeljko Medenica, Andrew L. Kun, 2007. *Comparing the Influence of Two User Interfaces for Mobile Radios on Driving Performance*, Driving Assessment, Stevenson, WA

## Biographical Sketch

Zeljko Medenica is a 2nd year PhD student at the Electrical and Computer Engineering department of the University of New Hampshire, under the supervision of Professor Andrew Kun. He obtained his BSc degree in Electrical and Computer Engineering at the University of Novi Sad in Serbia. His research interests include spoken human-computer interaction in mobile environments. Besides research, his interests include traveling, cycling, martial arts, and languages.

**Gregory J. Mills**

Interaction, Media and Communication Group
Department of Computer Science
Queen Mary, University of London.

gj@dcs.qmul.ac.uk

## 1 Research Interests

My research is primarily concerned with the empirical investigation of **human-human communication**, focusing on the interactive mechanisms deployed by interlocutors in **semantic co-ordination**. I am particularly interested in the mechanisms used for dealing with problematic understanding, e.g. **repair, clarification requests, corrections**, and their role in **language change** that occurs during the development of co-ordination. I am currently focusing on how **alignment** is exploited by interlocutors and also on the boundary between communication/**miscommunication**.

### 1.1 Past and Current Research

Perhaps one of the least contentious statements concerning language is that it is intrinsically underdetermined, dynamic, and adaptable to novel dialogue contexts. Despite this insight, research on dialogue systems has primarily focused on the information-exchange aspects of language use, pre-supposing that both interlocutor and dialogue manager already "know how to talk" about the particular domain and have already co-ordinated their linguistic resources to suit the communicative situation. Consequently, dialogue system implementations have traditionally relegated the importance of this co-ordination of linguistic resources by selective incorporation of highly domain-specific vocabularies and ontologies (Larsson, 2006).

The point of departure of my research is that this static treatment of language presupposes semantic transparency between interlocutors and is inadequate for describing how co-ordination is achieved. The inherently dynamic and adaptable nature of language, and the fact that interlocutors necessarily have different interaction histories suggests the importance of an account that explains how interlocutors manage to resolve these differences and converge on a semantic model during dialogue.

Although existing models of dialogue agree that this is achieved within interaction, they disagree on which particular interactive mechanisms are implicated in this process. The collaborative model of Clark et al (Clark, 1996) characterizes this as occurring through iterative cycles of "positive evidence of understanding", while the interactive alignment model (Garrod et al, 1994, 2004) prioritizes the role of priming in the development of co-ordination.

To address these differences and investigate in detail the negotiation of semantic co-ordination, my research involved a series of experiments using a novel text-based chat tool. In contrast to existing experimental approaches which involve relatively coarse modification of the communicative context, e.g. confederates or wizard-of-oz scenarios, the chat tool allows fine-grained manipulation of the unfolding dialogue by selectively interfering with individual co-ordination mechanisms (Healey and Mills, 2007).

Key findings from the experiments carried out with this chat tool are: interfering with sequential coherence of the dialogue leads participants to align more (Mills, 2007); interlocutors index the level of semantic co-ordination of partners with different levels of participation (Healey and Mills, 2006); and on encountering problematic utterances interlocutors resort to less specific and "vaguer" descriptions (Mills and Healey, 2006).

The findings from these experiments present co-ordination phenomena that are difficult to reconcile with both models, due to their "semantic neutrality" (Healey and Mills, 2006). In particular, the findings demonstrate how interlocutors are sensitive to semantic differences between different kinds of referring expression, and exploit these differences in order to develop and sustain mutual-intelligibility. Importantly, this is achieved tacitly by interlocutors, in particular when dealing with problematic understanding (Mills, 2007).

These findings also work against semantic transparency underwriting co-ordination. Interlocutors frequently introduce and use terms without having full understanding of their applicability to the dialogue situation. Instead, terms are introduced opportunistically, their meaning fleshed out through iterative cycles of repair (Healey and Mills, 2006).

My research has yielded rich patterns of co-ordination in clarification subdialogues that exhibit semantic change which are not strictly reducible to the exchange of propositionally encoded information, yet still have the effect of resolving problematic understanding. Importantly, they provide compelling evidence supporting the thesis that alignment is not simply an outcome of successful interaction, but is also exploited by interlocutors when dealing with problematic understanding (Mills, 2007; Mills and Healey 2008).

### 1.2 Current and future work

I am currently working on refining the experimental chat tool methodology, developing it as a general experimental platform for dialogue researchers, in order to facilitate the experimental investigation of dialogue phenomena[1].

I am also conducting experiments using this chat tool to investigate how interlocutors exploit alignment in dialogue, in particular on the role of figure and ground. I am also using the chat tool to investigate how these findings scale up to multilogue.

## 2 Future of Spoken Dialog Research

It is essential that the development of more naturalistic dialogue systems be guided by "what interlocutors actually do". All too frequently, the natural spontaneity, and flexibility of language and the interactive mechanisms used to deal with problems of mutual-intelligibility that arise when language is adapted to novel contexts of use is treated as inferior to the idealization of well-formed speech acts. This of prime importance, as experimental evidence demonstrates that repair facilitates comprehension (Brennan and Schober, 2001).

Further, this flexibility introduces the notion of semantic change occurring during the course of individual conversations, bringing the problem of semantic opacity to the foreground. Dialogue research could benefit from a critical assessment of the insights from the philosophy of language concerning intentional and also semantic transparency, in particular Wittgenstein's consideration of language as practice and his arguments concerning the limitations of a strictly informational view of language. This would address the issue that using a term necessarily introduces change into its meaning for an agent and the language community as a whole. Creating a typology of these changes, and how they are achieved through co-ordination mechanisms would decrease the gulf that exists between dialogues with dialogue managers and actual human-human conversation.

## 3 Suggestions for discussion

- Methodologies used to determine which particular mechanisms to include in dialogue systems: empirical approaches, e.g. wizard-of-oz, confederates, introduce different biases from user-simulations. Discussion of the relative merits and limitations of both would be of great benefit to dialogue system designers.
- Semantic and intentional transparency: what steps can be taken to make models of agency and language production/comprehension more similar to actual human-human conversation.
- Language change: how can dialogue managers be designed to allow for novel uses of terms by both interlocutors and dialogue system?

[1] DiET: Dialogue Experimentation Toolkit. http://www.dcs.qmul.ac.uk/research/imc/diet/

## References

Brennan, S.E., & Schober M. F. (2001) How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language, 44*, 274-296.

Clark, H. H. *Using Language*. Cambridge University Press, Cambridge.

Garrod, S. and Doherty, G. 1994. Conversation, co-ordination and convention: an empirical investigation of how groups establish linguistic conventions. *Cognition*, 53:181-215.

Healey, P.G.T. and Mills, G.J. 2006. "Participation, Precedence and Co-ordination in Dialogue" in *Proceedings of the 28th Conference of the Cognitive Science Society*. Vancouver, Canada. pp. 1470-1475.

Larsson, S. 2006. Semantic plasticity. Paper presented at LCM 2006. (Language, Culture and Mind), July 2006, Paris, France.

Mills, G. J. 2007. The development of semantic co-ordination in dialogue: the role of direct interaction. Unpublished PhD. Thesis.

Mills, G.J. and Healey, P.G.T. (2006) Clarifying Spatial Descriptions: Local and Global Effects on Semantic Co-ordination. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue*. University of Potsdam, Germany; pp.122-129.

## Biographical Sketch

The author completed his PhD in 2007 and is currently working as a Postdoc at Queen Mary University, focusing on the interactive mechanisms involved in semantic co-ordination. He is interested in the philosophy of language and mind, phenomenology of (mis)communication, the evolution of language, and also language use in therapeutic settings. Outside of academia, he flies gliders and plays electronic music and the harmonica.

**Crystal Nakatsu**

Department of Linguistics
The Ohio State University
1712 Neil Avenue
Columbus, OH 43201

cnakatsu@ling.osu.edu
linguistics.osu.edu/~cnakatsu

## 1 Research Interests

My research interests lie in the **natural language generation** aspect of **spoken dialogue systems**. I am interested in **generation methods** that can combine **statistical and symbolic methods** such as the **(over)generate-and-select** method. My current research focuses on using this method to appropriately generate **contrastive discourse connectives** in **comparative discourse**.

### 1.1 Ranking Realizations by Synthesis Quality

Following the approach by Walker et. al (2002), we utilized the generate-and-select method to select the best candidate realization from a set of synthesized realizations. We sought to determine whether we could train a ranker to select paraphrases that are predicted to sound natural when synthesized (Nakatsu and White, 2006).

First we generated a set of candidate realizations from a single disjunctive logical form, using OpenCCG (White, 2004; White, 2006). Then, using Festival (Taylor et al., 1998; Clark et al. 2004), we synthesized the set of candidate realizations, and had human judges rate the synthesized realizations for naturalness. Next we extracted features from the realizer (e.g. n-grams) and from the synthesizer (e.g. the join cost of the phonemes being synthesized). The ranker was then trained on different sets of these features along with the human ratings in a supervised fashion.

We tested the different rankers in a 10-fold cross-validation study. What we found was that the simplest ranker that only considered ngrams as features was the best predictor of the judges' score. Furthermore, this simple n-gram ranker performed similarly to the full-featured ranker that included target and join costs, and significantly better than a ranker that only considered target and join cost.

### 1.2 Error Analysis of Comparative Realizations

In recent work, Walker et. al. (2007) has applied the generate-and-select method to the restaurant comparison domain, and has included the use of discourse connectives (a.k.a. cue words) to overtly realize the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) discourse relations. Because their ranker does not exactly emulate the human ratings in its ranking of the candidate realizations, I have been studying their corpus and conducting an error analysis on the SPaRKy Restaurant corpus of generated restaurant comparisons (Walker et. al. 2007). Specifically, I'm interested in determining if there is a linguistic basis for the impressionistic human ratings, and if that linguistic basis can be exploited in the form of addition features for the ranker to learn.

What I find is that the human judges' seem to disprefer the use of multiple contrastive connectives (e.g. *however, on the other hand, while,* and *but*) in a single utterance (Nakatsu, 2008). Unsurprisingly, the ranker learns this preference. However, upon further analysis, it becomes apparent that the human judges do not unilaterally dislike multiple contrastive connectives, but rather that they are sensitive to the order of the arguments for certain contrastive connectives. This ordering constraint seems not to be captured by the ngram features. In addition, there are many more examples of the dispreferred usage of connectives in the corpus than of the preferred usage, leading the ranker to believe that use of connectives should generally be avoided.

Thus in future work, I plan to extract features which will hopefully capture this linguistic distinction. As well, I intend to investigate other linguistic factors that may explain the usage of these contrastive connectives, and attempt to capture these observations as features. This should increase the sensitivity of the ranker to the human judges' preferences for the use of these contrastive connectives.

## 2 Future of Spoken Dialog Research

In order for dialogue systems to become useful and-widespread, we need users to be successful in interacting with the system without extensive training. One stumbling block that seems to hinder users' adoption of the spoken dialogue systems is that they have trouble gauging what the system is able to comprehend, given that the output of dialogue systems is often very natural sounding, but not all input is recognized. Apart from solving the knowledge problem of NLU, future work on dialogue systems should include research on which verbal (and/or non-verbal) cues help users to discern the language competency of other humans, and how we can generate those cues so that users can quickly discern the verbal capabilities of the automated spoken dialogue system.

## 3 Suggestions for discussion

- Statistical methods: Have we hit the ceiling in terms of what statistical modeling is capable of capturing? Do we need to start augmenting statistical language modeling with linguistic theory, or do we just need to find a better algorithm?
- User expectations: Would systems that generated natural sounding output create false expectations for users, who may then become frustrated with the system's inability to converse with the human? Would it instead be better to match the level of generated output with the level of system capability (i.e. give some verbal cues in the generated speech that indicates a limited communicative ability, such as childlike/non-native-like speech for simple dialog applications).
- User Cognitive Load: Structuring the delivery of information eases the cognitive strain on the user of an information delivery system. How should information be structured, how much information should be given to the user in a single turn, and how complex should sentences in the turn be in order for the user to retain the information being given? Does it depend what modality is being used to disseminate the information? What is the impact on the cognitive load of using similar, but not quite synonymous discourse connectives to help structure the information?

## References

Robert A.J. Clark, Korin Richmond, and Simon King. 2004. Festival 2 – build your own general purpose unit selection speech synthesiser. In *5th ISCA Speech Synthesis Workshop*, pages 173–178, Pittsburgh, PA.

Nakatsu, C., & White, M. (2006). Learning to say it well: Reranking realizations by predicted synthesis quality. In Proc. of the Annual Meeting of the Association for Computational Linguistics

Nakatsu, C. (2008). Learning Contrastive Connectives in Sentence Realization Ranking. In Proc. of the 9th SIGDial Workshop on Discourse and Dialogue. To appear.

William C. Mann and Sandra A. Thompson. 1987. Rhetorical structure theory: A theory of text organization. Technical Report ISI/RS-87-190, University of Southern California Information Sciences Instuture.

P. Taylor, A. Black, and R. Caley. 1998. The architecture of the the Festival speech synthesis system. In *Third International Workshop on Speech Synthesis, Sydney, Australia*

Marilyn A.Walker, Owen C. Rambow, and Monica Rogati. 2002. Training a sentence planner for spoken dialogue using boosting. *Computer Speech and Language*, 16:409–433.

M. Walker, A. Stent, F. Mairesse, and Rashmi Prasad. 2007. Individual and domain adaptation in sentence planning for dialogue. Journal of Artificial Intelligence Research (JAIR), 30:413–456.

Michael White. 2004. Reining in CCG Chart Realization. In *Proc. INLG-04*.

Michael White. 2006. CCG chart realization from disjunctive logical forms. In *Proc. INLG-06*.

## Biographical Sketch

Crystal Nakatsu received her B.A. in Linguistics from the University of British Columbia in Vancouver, Canada, where she was born and raised. She is currently a Ph.D. student in the Linguistics department of The Ohio State University under the supervision of Dr. Michael White, and studies natural language generation for use in spoken dialogue systems. Throughout her undergraduate and graduate career, Crystal has held internships at a number of companies and institutions, including Webmind, Inc. (formerly Intelligenesis, Inc.), BCL Technologies, the Spoken Dialog Group @ NASA Ames, and Articulab @ Northwestern University.

**Oskar Palinko**

Electrical and Computer Engineering Dept.
University of New Hampshire
Kingsbury Hall
Durham, New Hampshire, USA

oskar.palinko@unh.edu

## 1 Research Interests

My broad field of research includes **automotive spoken language systems** and **multimodal ubiquitous systems**. Specifically, I am interested in researching different **modalities of push-to-talk speech user interface activation** and how these systems affect driving and spoken language system usage in vehicles. I am also investigating how different wearable push-to-talk solutions (e.g. glove) could create a **continuous, ubiquitous speech user interface** experience in and outside a vehicle.

### 1.1 Push-to-talk glove

In a recent study (Kun et al., 2007), our group has found that push-to-talk usage, which is one characteristic of the Project54 in-car speech user interface (Kun et al., 2004), influences driving performance. This motivated us to explore other possible push-to-talk (PTT) solutions, which could have a lesser impact on driving. Therefore I proposed a wireless PTT glove solution, which allows the user to activate the Project54 speech user interface in a less restrictive way compared to the fixed PTT. We performed a pilot experiment with this PTT solution in our high-fidelity driving simulator (Palinko and Kun, 2008a). The subjects drove a city scenario, which consisted of intersections, straight portions and curvy roads. The results show that when using the glove, participants tended to operate the PTT in positions other than that provided by the fixed PTT button, which is positioned on the crossbars of the steering wheel. The glove also significantly reduced reaction times of operating the PTT, while taking turns at intersections.

### 1.2 Steering wheel sensor

In order to provide an even less restrictive and more transparent PTT activation method, we integrated pressure sensor strips onto the perimeter of the steering wheel. In our paper describing this system (Palinko and Kun, 2008b), we hypothesize that a steering wheel could provide an effective PTT sensing surface without a negative influence on driving performance. In our experiments, participants drove in a city scenario, while activating the speech recognizer by double-tapping on the steering wheel sensor. This solution was compared to the standard fixed PTT activation method. The results show, that the new system does not degrade the driving performance of subjects as compared to using the fixed PTT button. In post-experiment interviews, the participants reported finding the steering wheel sensor to be an interesting, useful and convenient method of PTT activation.

### 1.3 Push-hold-release vs. push-release

I am currently conducting an experiment with 20 subjects which will evaluate how different aspects of in-car speech user interface design influences driving performance. The experiment focuses on speech recognizer accuracy, PTT activation method (glove vs. fixed), and PTT activation sequence. I am investigating two activation sequences: push-hold-release and push-release. In the first one, subjects push and hold down the PTT button while issuing commands, while in the second one they just push and release the PTT indicating the starting point of an uterance. In this case the end point is automatically detected. We hypothesize that the latter approach will have a beneficial effect on driving performance, since the push-button does not have to be held down for a longer period of time (1-2 seconds). The holding action may interfere with driving in curves or while taking turns, since the driver’s hand must be held at a constant position on the steering wheel for some time.

### 1.4 Ubiquitous computing with the PTT glove

In the near future, we plan to transform our in-car wireless PTT glove into a ubiquitous input device that first responders (mainly police officers) could use inside and outside their vehicles. The PTT glove will become the PTT solution for our handheld system, which connects directly to the Project54 base system inside the vehicle. This will allow officers to issue speech commands to the handheld, without actually having to push buttons on the device. The speech commands will be relayed to the in-car system. The current version of our handheld application uses the handheld's built-in buttons for PTT activation and it is discussed in a recent paper by our group (Fekete and Kun, 2008).

## 2 Future of Spoken Dialog Research

In my opinion spoken dialog research should address the question of providing a more natural speech dialog environment. In such a system, the user would feel more comfortable talking to a spoken dialog system, since the machine would successfully mimic an intelligent speaking partner: it should be able to detect different emotions in the speaker's voice and it should provide appropriate prosody in its voice, even show compassion or share a laugh. Much attention should be focused on avoiding misinterpretation of the user's emotional status. In the case of a misinterpretation, the users could feel that the spoken dialog system is trying to mock their feelings and might consider it to be an unworthy speaking partner (Nass and Brave, 2005).

In order to be able to develop such semi-intelligent spoken dialog systems, it is necessary to deepen our understanding of human-human speech interaction. One of the steps in that direction is to be able to detect and interpret multi-threaded speech dialogs (Shyrokov et al., 2007). Another obvious need is to detect and analyze emotions in human-human and human-computer conversations, e.g. by applying speech processing techniques in order to analyze prosody.

## 3 Suggestions for discussion

Proposed discussion topics:

- Emotionally aware spoken dialog systems. How to get there?
- Need for push-to-talk signaling for speech interaction in noisy conditions (e.g. in-car) with multiple speech sources (e.g. driver and multiple passengers).
- Speech interaction with vehicles. Should cars be perceived by drivers as semi-intelligent or should the vehicle not display signs of situation awareness using speech?

## References

Andras K. Fekete and Andrew L. Kun. 2008. To appear. *Handheld Computing in Law Enforcement: A Pilot Study*. Fourth IET International Conference on Intelligent Environments (IE08), Seattle, WA, July 21-22, 2008.

Andrew L. Kun, W. Thomas Miller, III and William H. Lenharth. 2004. *Computers in police cruisers*. IEEE Pervasive Computing, Volume 3, Issue 4, Pages: 34 - 41.

Andrew L. Kun, Tim Paek, Zeljko Medenica. 2007. *The Effect of Speech User Interface Accuracy on Driving Performance,* Interspeech 2007, Antwerp, Belgium.

Clifford Nass and Scott Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Interaction*. MIT Press, Cambridge, MA.

Oskar Palinko and Andrew L. Kun. 2008a. To appear. *Prototype Wireless Push-To-Talk Glove.* Fourth IET International Conference on Intelligent Environments (IE08), Seattle, WA, July 21-22, 2008.

Oskar Palinko and Andrew L. Kun. 2008b. To appear. *Steering Wheel Sensor as a Push-To-Talk Solution.* Fourth IET International Conference on Intelligent Environments (IE08), Seattle, WA, July 21-22, 2008.

Alexander Shyrokov, Andrew L. Kun and Peter Heeman. 2007. *Experimental Modeling of Human-Human Multi-Threaded Dialogues in the Presence of a Manual-Visual Task.* SIGdial 2007, Antwerp, Belgium, September 1-2, 2007.

## Biographical Sketch

Oskar Palinko is currently a 2nd year Master's student and Research Assistant at the Electrical and Computer Engineering Department of the University of New Hampshire. He works under the supervision of Professor Andrew L. Kun. He earned his BSc in Electrical and Computer Engineering at the University of Novi Sad, Serbia in 2004. After his graduation he worked as a Teaching Assistant at the same university in the field of Robotics and Automation. His extra-curricular interests include astronomy, zen, web design, kayaking and mountain biking.

**Joana Paulo Pardal**

Dept. Computer Science and Engineering
IST, Technical University of Lisbon
L$^2$F, Spoken Language Systems Laboratory
R. Alves Redol, 9 - 2 Esq - sala 234A
1000-029 Lisboa, Portugal

joana@l2f.inesc-id.pt
http://www.l2f.inesc-id.pt/˜joana

# 1 Research Interests

My research interests lie generally in the area of **spoken dialogue systems** with particular interest in **software engineering** techniques to **dynamically integrate structured knowledge sources**, like databases and **ontologies** (Paulo Pardal, 2007), and in **evaluation frameworks** that allow measuring the advantages. The challenges of creating **tutorial and educational systems** that can be used by the **general public** at their homes, schools or museums are also part of my research. This includes the use of general resources like those usually provided by Question-Answering systems (Mendes, 2008). Finally, I am also interested in the creation of emotion models to enrich the agents with **emotional behavior** that changes according to the dialogue's flow and the systems' success.

## 1.1 Background and related work

DIGA (DIaloG Assistant), the domain-independent framework for Spoken Dialogue Systems of L$^2$F was created back in 2000 (Mourão et al., 2004) The framework is highly inspired on the TRIPS architecture (Allen et al., 2005): it is frame-based and is used to build domain-specific dialogue systems. Every domain is described by a frame, composed by domain slots that are filled with user requests. Several systems were already created with this framework: a bus ticket vending system, that provides access to buses timetables; a digital virtual butler named *Ambrósio* that controls home devices; a prototype system that helps the user perform some task (tested for cooking and automobile repair domains); and two telephone-based systems (home banking and digital personal assistant). *Ambrósio* is publicly available at the "House of the Future" (www.casadofuturo.org), on the Portuguese Telecommunications Museum since 2003. The telephone-based systems were developed under the project TecnoVoz (www.tecnovoz.pt), a Portuguese national consortium including Academia and Industry partners. The integration of the framework into commercial products led to a major reengineering process (Martins et al., 2008b). Also, some improvements on the parsing method were needed to deal with different data types (Martins et al., 2008a).

## 1.2 Past and current work

Most practical dialogue systems are designed for a specific task, and even if the authors were concerned with possible future extensions, integrating new tasks is always a challenge. Work has been done to take advantage of some programming paradigms to ease this process. Dynamic integration of new tasks according to some kind of structured knowledge is an interesting research topic.

The use of databases has been shown to ease the extension of a system to new tasks since we can extract domain knowledge from the tables' structure and create systems that generically use that kind of information. Given that the ontologies can be seen as an upgrade of databases (whereas richer information can be stored) a new methodology will be proposed that will use domain knowledge collected in a ontology, gathering the (currently scattered) domain knowledge into a specific module. This will help when introducing new languages in the system; it also pushes the dialogue phenomena into a specific module that can be reused across different systems. Better module APIs are needed to do so.

To test this possibility, an ontology for the cooking domain concepts was built (Ribeiro et al., 2006) that was later populated with information automatically extracted from books and web sites with a natural language specific tool (Machado, 2007). Also, a first prototype was built that helps the user with the tasks needed to perform a chosen recipe.

Current work includes a survey of the state of the art and a technical report summarizing the existing dialogue systems (with a categorization according to the tasks that are performed and to the information that is used); and the systems using ontologies. Also a state of the art on generic systems that takes advantage of the use of ontologies (against the sole use of databases).

## 1.3 Future work

After DIGA has been tested as a simple tutorial system (where the system, instead of receiving orders, instructs the user to perform some steps towards a selected task), I will adapt the existing module to take advantage of ontological knowledge. The current version loads recipes from a basic XML file. The next version will obtain that information from the database populated according to the cooking ontology. Later I will include the ontological knowledge in the system and will measure the impact of gathering the knowledge in a single data source. It is my belief that we can extract all the domain-specific knowledge that needs to be included in the systems' modules from a domain-specific ontology (vocabulary to recognize and use, domain terminology and thesaurus, translations between natural language and domain-specific keywords, etc.). It will also be interesting to declare new dialogue systems as instances of an ontology stating the needed information to declare a system. To allow real-time responses to the users it might be necessary to create some optimization mechanisms that process the knowledge stored in the ontology offline in order to enhance the integration at runtime.

## 2 Future of Spoken Dialog Research

Currently spoken dialogue systems are proposed only when no other input modalities are available (like when there is no access to a keyboard or when the user has some kind of special need – blind, reduced accessibility, etc.) However, we should consider the use of speech whenever it is natural. That would be easier if interaction with this systems was more natural (more similar to human-human interaction).

Research in psycholinguistics has shown that continuous understanding plays a major role in language understanding by humans. This can be seen by the completion of what the interlocutor is saying or by response earlier than the end of a sentence. Capturing human continuous understanding behavior in a multimodal dialogue corpus (Gomez-Gallo et al., 2007) is an initial step towards natural interaction.

When human-computer interaction approaches human-human interaction, people will feel comfortable on delegating some tasks to a digital helper while they will concern themselves with some other tasks. This also needs to consider the right time to interrupt and managing priorities.

## 3 Suggestions for Discussion

- Question-Answering: architectural needs and dialogue handling to take advantage of the knowledge store in the database/ontology.
- Teaching SDS (methods, frameworks, evaluation) and SDS for Teaching (tutorial and educational applications).
- Evaluation: universal metrics for comparing disparate systems, tasks, languages and modalities; expert systems against rapid development frameworks.

## References

J. F. Allen, G. Ferguson, M. Swift, A. Stent, S. Stoness, L. Galescu, N. Chambers, E. Campana, and G. Aist. 2005. Two diverse systems built using generic components for spoken dialogue (recent progress on TRIPS). In *Proc. of the Interactive Poster and Demonstration Sessions - ACL.*

C. Gomez-Gallo, G. Aist, J. F. Allen, W. de Beaumont, S. Coria, W. Gegg-Harrison, J. Paulo Pardal, and M. Swift. 2007. Annotating continuous understanding in a multimodal dialogue corpus. In *SEMDIAL – DECALOG.*

T. Machado. 2007. Extracção de informação – introdução automática de receitas de acordo com ontologia. Master's thesis, IST, UTL.

F. Martins, A. Mendes, J. Paulo Pardal, N. J. Mamede, and J. P. Neto. 2008a. Using system expectations to manage user interactions. In *PROPOR 2008 (to appear)*, LNCS. Springer.

F. Martins, A. Mendes, M. Viveiros, J. Paulo Pardal, P. Arez, N. J. Mamede, and J. P. Neto. 2008b. Reengineering a domain-independent framework for spoken dialogue systems. In *Proc. software eng., testing, and quality assurance for NLP (to appear).* ACL WS.

A. Mendes. 2008. Introducing dialogue in a QA system. In *Doctoral Symposium of 13$^{th}$ Intl. Conf. Apps. Nat. Lang. to Information Systems, NLDB (to appear).*

M. Mourão, R. Cassaca, and N. J. Mamede. 2004. An independent domain dialogue system through a service manager. In *EsTAL*, LNCS. Springer.

J. Paulo Pardal. 2007. Dynamic use of ontologies in dialogue systems. In *NAACL-HLT Doctoral Consortium.*

R. Ribeiro, F. Batista, J. Paulo Pardal, N. J. Mamede, and H. S. Pinto. 2006. Cooking an ontology. In *AIMSA*, LNCS. Springer.

## Biographical Sketch

Joana Paulo Pardal is a 3$^{rd}$ year Ph.D. student in Informatics and Computer Science Engineering at IST, UTL, under the supervision of Nuno J. Mamede (IST), H. Sofia Pinto (IST) and James F. Allen (U. Rochester). She holds a fellowship from FCT (Portuguese Nat. Science Foundation). Joana received a *licenciatura* (a 5 year full-time degree) in 2001, and an M.Sc. in 2004 both in Informatics and CS Eng. from IST. Joana is a researcher at L$^2$F since 2001. She is a Lecturer at IST since 2002. She is a student member of ISCA, ACL and AAAI and participates on CMU *DoD* reading group. In 2004 she worked at GRIL, U. Blaise Pascal, Clermont-Ferrand, France. In 2006 she spent a research term at the CS Dept., U. of Rochester, NY, USA.

**David Díaz Pardo de Vera**

Universidad Politécnica de Madrid
Cuidad Universitaria s/n
28040 Madrid
Spain

dpardo@gaps.ssr.upm.es

## 1 Research Interests

My main research interests lie in **spoken dialogue systems** (SDSs) incorporating **nonverbal channels of communication**, focussing particularly on **user-centred quality evaluation**.

### 1.1 Past and ongoing research: ECAs

At GAPS (the Signal Processing Applications Group at UPM) we have been studying a variety of contexts of multimodal human-machine interaction in which dialogue systems are very much at the core. Our main goal is to improve the robustness of human-machine communication while making the experience for the users as efficient and enjoyable as possible.

We have developed several interaction platforms to test various modalities of interaction for a variety of purposes. One such platform enabled us to compare several forms of biometric access (fingerprint, signature, voice and iris recognition), using either only text prompts to guide the user, or spoken messages accompanied by an avatar. We found that, overall, the presence of an avatar tended to enhance the feeling of pleasantness and the perception of ease of use. However, the concern over privacy was intensified.

In another study we sought to find the effects that incorporating an embodied conversational agent (ECA) might have on the usability, efficiency and user acceptance of an identity verification system relying on voice recognition. In this case we have a spoken dialogue interface which incorporates an animated figure. Again, interaction with an animated figure was more efficient and enjoyable, but it also enhanced users' privacy concerns (Hernández et al., 2007).

We then tested an application to remotely control home devices (i.e., a domotic system). Again, we compared two spoken dialogue interfaces: one with and the other without a gesturing ECA onscreen. In addition to the gestures designed to show the state of the dialogue and the system's level of recognition confidence, and also those to keep users in a positive frame of mind when recognition errors occurred, we tried out a proxemic code to mark dialogue turns. Our gestural and proxemic cues helped users in their interactions with the system, making these more fluent, and the users more confident. The error management strategies we tried (e.g., implicit confirmation of information, apologising and showing special interest in getting it right in the following attempt) seem to work better when accompanied by appropriate gestures. On the other hand, we did also found mild evidence that an ECA can cause users to initially overestimate the system's conversational capability, leading to some degree of disappointment after the interaction.

### 1.2 Past and ongoing research: evaluation

A related area of research we have been working on is dialogue system evaluation. We may identify two main goals of user-centred evaluation: the first is to be able to establish the users' overall opinion of the "quality" of a dialogue system as a whole, and possibly of a variety of system characteristics (such as intelligibility, quality of recognition, fluency or "cognitive" ability); the second is to develop models with which the users' perception of quality may be predicted from a set of parameters. We have chosen to focus on the facet of quality that has to do with user acceptance, which in turn has elements of user satisfaction and inclination to use the system.

Various evaluation schemes have been proposed in the literature, but as far as we know there is no standard procedure to evaluate dialogue systems in specific contexts, such as enrolment and verification dialogues for speaker authentication systems, which, as we mentioned earlier, is an area of research we have payed special attention to. Nor have we found well defined and tested guidelines to evaluate multimodal dialogue systems that include a humanlike figure in the visual communication channel (an ECA), or standardised approaches to take user emotions into account.

What we have done in order to incorporate these aspects in our evaluations is to follow the ITU P.851 recommendation (ITU-T P.851, 1999) on questionnaire design for the evaluation of spoken dialogue systems for general telephone services, and expand it to include dimensions (as sets of questions) to evaluate user perceptions related with secure access and with the ECA. Inspired by Möller's taxonomy of quality factors (Möller et al., 2007), in our approach we combine user's responses to questionnaires with performance and interaction data registered automatically.

Rather than establish a rigid taxonomy of parameters that may be related to user acceptance (understood as an overall aspect of quality), we have defined a quality-parameter class structure, or frame (since we hope it will be useful as a "frame" for modelling), that provides a certain conceptual clarity both to further develop questionnaires and future experiments and to analyse the data obtained in them. The class structure arises from considering two orthogonal sets of dimensions. First, from the literature on usability and ergonomics (among which Angela Sasse's work is especially relevant to ours; see, e.g., Sasse (2004)) we have extrapolated the notion that three major classes of parameters are generally (and implicitly) considered to be related to user acceptance: likeability factors (those that have to do with the experience of using the system), rejection factors (those that can only have a negative valence) and perception of usefulness (those related with how well the user believes the system is suited to the pursuit of the goals she would expect or want to achieve by using it). Secondly, the class structure is further broken down into various levels that may be independently subject to, or may independently affect, user perception. We may distinguish an overall system-assesment level, task and goal-related levels (here we might have the biometric access task and the accessed application), and an interaction-through-interface level (e.g., a dialogue system with or without an ECA).

## 2 Future of Spoken Dialog Research

I believe in the near future we will see an intensification of efforts to standardise the specification of communication acts –incorporating speech and visual communication (e.g., ECA gestures)– for system dialogue output. The communication act generation process needs to be studied, in the first place to determine what stages it should be broken down into. We need to know how the system should react to user input as a function of the interaction context, how to specify multi-layered and multimodal semantic output, from a pre-verbal, conceptual stage to a verbal, message and gesture-bound one. Efforts from the AI camp centred on cognition are already taking off (e.g., the SAIBA framework). I believe an HMI counterpart focusing on the *interaction* is needed.

Multi-layered semantic appraisal of the user's messages and behaviour may possibly take longer, but is nonetheless also necessary in order to fully achieve a human interface metaphor.

## 3 Suggestions for discussion

My suggestions for discussion are the following:

- What elements of user-centred quality should we focus on when evaluating multimodal dialogue systems? Identifying factors that affect user perception of quality and how to determine the relevant interrelations.
- Formalisation of multimodal communication act generation.
- Bridging the gap between the AI and the HMI approaches to dialogue generation.

## References

A. Hernández, B. López, D. Díaz, R. Fernández, L. Hernández, and J. Caminero. 2007, *A person in the interface: effects on user perceptions of multibiometrics*, Workshop on Embodied Language Processing, in the 45th Annual Meeting of the Association for Computational Linguistics, ACL, pp. 33-40, Prague.

ITU-T P.851. 1999, *Subjective Quality Evaluation of Telephone Services Based on Spoke. Dialogue Systems*, International Telecommunication Union (ITU), Geneva.

S. Möller, P. Smeele, H. Boland, and J. Krebber. 2007, *Evaluating spoken dialogue systems according to de-facto standards: A case study,* Computer Speech & Language 21 26-53.

SAIBA: http://wiki.mindmakers.org/projects:saiba:main/

M. A. Sasse. 2004. *Usability and trust in information systems*, Cyber Trust & Crime Prevention Project. University College London.

## Biographical Sketch

David Díaz Pardo de Vera has an MEng in telecommunications engineering from Universidad Politécnica de Madrid. He is currently a 1$^{st}$ year PhD student and holds a research position in HMI in the same university, under the supervision of Luis Hernández. David is also writing a masters' thesis on the sociology and ethics of human-animal relations in the context of tourism in Lapland and in Spain.

# Antoine Raux

Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
USA

antoine@cs.cmu.edu
www.cs.cmu.edu/~antoine

## 1 Research Interests

I am interested in all aspects of practical spoken dialog systems, from **speaker adaptation** for ASR, to **prosody modeling** for TTS, to **dialog management** and, most importantly, **turn-taking**. The main purpose of my work is to improve human-machine spoken interaction by enhancing the low level abilities (ASR, TTS, channel establishment) of spoken dialog systems.

### 1.1 Previous work

*Prosody modeling for speech synthesis*
Originally as part of a class project on TTS, I investigated ways to create more appropriate and natural prosodic contours for dialog utterances. In (Raux and Black, 2003), we describe a new method to generate F0 contours by concatenating portions of natural contours from recorded utterances, in the same way that concatenative speech synthesis generates utterances by concatenating portions of waveforms. This approach allowed us to build natural yet flexible contours in a very limited amount of time (assuming a database of the target phenomenon is available) and without requiring an expert to write prosodic rules.

*Non-native users of dialog systems*
Another of my topics of research has been non-native speakers, and how to improve their experience with spoken dialog systems. From data collected with the Let's Go system, I studied linguistic (Raux and Eskenazi, 2004), phonetic (Raux, 2004), and acoustic differences between native and non-native users, and proposed a method to guide non-native users' language towards the lexical and syntactic structures expected by the system through lexical entrainment.

### 1.2 Thesis work

*Let’s Go!*
Most of my empirical research is based on the Let’s Go! bus information system. Built and maintained in collaboration with Brian Langner, under the supervision of Alan Black and Maxine Eskenazi, Let’s Go! is a telephone-based dialog system used by the general Pittsburgh population (the system is available at night, when the human operators of the bus company are not working). In addition to providing large amounts of data (40-80 dialogues a day on average), this system allows us to directly test the impact of research ideas on dialogs with real users. Interesting issues that we have faced include how much do users actually listen to what the system says, how they respond to different prompts, and how to handle non-understandings (Raux et al 2005, Raux et al 2006).

*Olympus*
Let’s Go! is based on a generic spoken dialog system architecture that has been created over the past 10 years at Carnegie Mellon, starting from the DARPA Communicator project. I joined Dan Bohus in his effort to turn the original Communicator system into a domain-independent platform for dialog systems research (Bohus et al, 2007). Our recent efforts, with Thomas Harris, have focused on making Olympus easier to distribute, install and use, including tutorials and documentation, with the idea of making it a true large scale open-source project supported by a broad community.

*Turn-Taking*
While most research in spoken dialog systems has focused on task and on dealing with uncertainty in speech recognition and understanding, very little attention has been given to the way the system and the user manage turn-taking. In the vast majority of cases, spoken dialog systems assume a rigid turn-taking behavior where user input goes through a serial understanding-dialog management-generation pipeline and turn boundaries are detected using only pause information. This leads to interaction issues like turn overtaking and unnecessary delays. My current research aims at addressing these issues by:

1) integrating an **interaction manager** into Olympus, in conjunction with the higher level, traditional dialog manager. I described this multi-layer architecture in (Raux and Eskenazi, 2007),
2) building models of **endpointing** and **interruptions** to better handle the crucial conver-

sational phenomena of *gaps* and *overlaps*. My SIGdial paper (Raux and Eskenazi, 2008) addresses the problem of endpointing, while I'm currently working on interruptions,

3) endowing the interaction manager with **reactive error handling** abilities in order to properly deal with floor conflicts. This is future work, enabled by the internal structure of the interaction manager as described in (Raux and Eskenazi, 2007).

## 2 Future of Spoken Dialog Research

With the advent and spread of practical application of speech technology, I think the research should diversify into different directions. Research that tries to make systems more human-like should be paralleled by research on designing speech and multimodal interfaces. While most people would agree on the intellectual merit of the former type of research, designing practical interfaces is an HCI issue that does not necessarily need to follow the human-human model. These two approaches to speech application research are sometimes contrasted in critiques or defenses of specific projects. Obviously, both views share common issues (e.g. ASR, turn-taking…) and cross-fertilization should happen. However, both would also benefit from a clear definition of their distinct goals. On the one hand, human-like agent research offers a long term grand challenge as well as potential application to specific tasks (e.g. games, tutoring) where the consensus is that being human-like is a key feature. On the other hand, there is still much room for research on new interaction paradigms for speech and, in particular, multimodal applications.

I hope to see progress on both fronts in the coming decade. The very young field of speech-based human-robot interaction is one direction for human-like agent research. It offers aspects like situatedness that were not part of previous, mostly telephone- or desktop-based systems. Mobile devices, where one can combine small touchscreen/keyboard and speech are a platform of choice for the second type of research.

## 3 Suggestions for discussion

*Realistic conversation*: what is missing to accomplish human-like conversation on simple/small domains? Better speech recognition? More control over language generation/speech synthesis? Smoother turn-taking?

*Embodied agents*: what are the potential applications of humanoid embodied agents (either on-screen or as solid robots)? Tutoring, gaming, caretaking, …? What kinds of dialog need to be supported for such applications (social, guidance, information access, …)?

*Mobile devices*: how can we efficiently combine a small display/touch screen and speech input/output to optimize user experience? What applications are best suited for such a multimodal approach?

## References

Antoine Raux and Alan W Black. *A Unit Selection Approach to F0 Modeling and its Application to Emphasis*, ASRU 2003.

Antoine Raux and Maxine Eskenazi. *Non-Native Users in the Let's Go Spoken Dialogue System: Dealing with Linguistic Mismatch*, HLT 2004.

Antoine Raux. *Automated Lexical Adaptation and Speaker Clustering based on Pronunciation Habits for Non-Native Speech Recognition*, ICSLP 2004.

Antoine Raux, Brian Langner, Dan Bohus, Alan Black and Maxine Eskenazi. *Let's Go Public! Taking a Spoken Dialog System to the Real World*, Interspeech 2005.

Antoine Raux, Dan Bohus, Brian Langner, Alan Black and Maxine Eskenazi. *Doing Research on a Deployed Spoken Dialogue System: One Year of Let's Go! Experience*, Interspeech 2006, Pittsburgh, USA.

Dan Bohus, Antoine Raux, Thomas Harris, Maxine Eskenazi and Alex Rudnicky. *Olympus: an open-source framework for conversational spoken language interface research*, HLT-NAACL 2007 workshop: Bridging the Gap in Dialog Technology.

Antoine Raux and Maxine Eskenazi. *A Multi-Layer Architecture for Semi-Synchronous Event-Driven Dialogue Management*, ASRU 2007.

Antoine Raux and Maxine Eskenazi. *Optimizing Endpointing Thresholds using Dialogue Features in a Spoken Dialogue System*, SIGdial 2008.

## Biographical Sketch

I am a PhD candidate at the Language Technologies Institute, Carnegie Mellon University. Prior to coming to CMU, I got an engineering degree from Ecole Polytechnique, France, and a Masters Degree from Kyoto University, Japan. For 6 years, I have been blessed with a wonderful wife Miyako, and we have two energetic children, Yuma, 4, and Manon, 1.

**Verena Rieser**

Human Communication Research Centre
Edinburgh University
vrieser@inf.ed.ac.uk
www.coli.uni-saarland.de/~vrieser/

## 1 Research Interests

My research interests lie in the area of statistical machine learning techniques and spoken dialogue systems. In particular, I am interested in using Reinforcement Learning (RL) to design strategies for Spoken Dialogue Systems (SDS), see (Young, 2000), (Lemon and Pietquin, 2007). One of the key advantages of statistical optimisation methods, such as RL, for dialogue strategy design is that the problem can be formulated as a principled mathematical model which can be automatically optimised by training on real data.

In my work I investigate the use of RL for different aspects of system design: Dialogue Management (DM) and Natural Language Generation (NLG).

### 1.1 Optimising Dialogue Strategies

For my Ph.D. I applied RL to optimise dialogue strategies for multimodal dialogue systems (Rieser, 2008), (Rieser and Lemon, 2008a), (Rieser and Lemon, 2008b). In particular, we propose to learn dialogue strategies by simulation-based RL, where the simulated environment is learned from small amounts of Wizard-of-Oz (WOZ) data. Using WOZ data rather than data from real Human-Computer Interaction (HCI) allows us to learn optimal strategies for new application areas beyond the scope of existing dialogue systems. Optimised learned strategies are then available from the first moment of online-operation, and tedious handcrafting of dialogue strategies is fully omitted. We call this method 'bootstrapping'.

We apply this framework to optimise multimodal information-seeking dialogue strategies for an in-car MP3 music player. Dialogue Management and multimodal output generation are two closely interrelated problems for information seeking dialogues: the decision of *when* to present information depends on *how many* pieces of information to present and the available options for *how* to present them, and vice versa. We therefore formulate the problem as a hierarchy of joint learning decisions which are optimised together. We see this as a first step towards an integrated statistical model of DM and output planning (NLG).

### 1.2 Optimising Output Generation

In current research we apply the introduced bootstrapping approach to optimise Natural Language Generation in spoken dialogue (see the EC FP7 CLASSiC project: www.classic-project.org). We will follow a similar overall aim: to improve the global user experience by optimising local dialogue decisions (Lemon, 2008). We will also explore RL for more complex domains such as troubleshooting dialogue (Janarthanam and Lemon, 2008).

## 2 Future of Spoken Dialog Research

One current vision for SDS research is to turn these systems into "organic interfaces" (Zue, 2007), i.e. interfaces which automatically adapt to the current context and the user through learning. However, statistical learning techniques are still not widely used. Major challenges to overcome is the need for training data, tractability of learning with large state spaces, learning with Partially Observable Markov Decision Processes (POMDPs), and quality assurance for simulated learning environments (Lemon and Pietquin, 2007).

Besides these technical challenges, statistical techniques, such as RL, also lack acceptance amongst the dialogue community. Especially their use for commercial application has been questioned (Paek, 2006). Research systems traditionally are logic-based, e.g. (Grosz and Sidner, 1986), (Steedman and Petrick, 2007), whereas industry (still) relies on simple Finite State Automaton (Pieraccini and Huerta, 2005). We believe statistical approaches to SDS design have the potential to "bridge the gap" between industry and research: one the one hand data-driven approaches provide insight in the general mechanisms underlying communication; on the other hand they facilitate rapid development of robust strategies.

## 3 Suggestions for Discussion

One feature which makes machine learning techniques to SDS particular attractive, is the fact that it drives SDS development towards being more scientific research. Dialogue design in the past was "more art than science" (Jones and Galliers, 1996). We therefore suggest the following topics for discussion:

- What does it take to make dialogue design proper scientific research?
- How can we facilitate comparison between different dialogue design approaches? Would a shared task

effort (as carried out for other areas of Computational Linguistics) be suitable for SDS design?

- How can we address the need for dialogue corpora?

## References

Grosz, B. J. and Sidner, C. L. (1986). Attention, intentions and the structure of discourse. *Computational Linguistics*, 12(3-4):175–204.

Janarthanam, S. and Lemon, O. (2008). User simulations for online adaptation and knowledge-alignment in troubleshooting dialogue systems. In *Proc. of the 12th SEMdial Workshop on on the Semantics and Pragmatics of Dialogues*.

Jones, K. and Galliers, J. (1996). *Evaluating Natural Language Processing Systems: An Analysis and Review*. Springer Verlag.

Lemon, O. (2008). Adaptive natural language generation in dialogue using Reinforcement Learning. In *Proc. of the 12th SEMdial Workshop on on the Semantics and Pragmatics of Dialogues*.

Lemon, O. and Pietquin, O. (2007). Machine Learning for spoken dialogue systems. In *Proc. of the International Conference of Spoken Language Processing (Interspeech/ICSLP)*.

Paek, T. (2006). Reinforcement Learning for spoken dialogue systems: Comparing strengths and weaknesses for practical deployment. In *Proc. Dialog-on-Dialog Workshop, Interspeech*.

Pieraccini, R. and Huerta, J. (2005). Where do we go from here? Research and commercial spoken dialog systems. In *Proc. of the 6th SIGdial Workshop on Discourse and Dialogue*.

Rieser, V. (2008). *Bootstrapping Reinforcement Learning-based Dialogue Strategies from Wizard-of-Oz data (to appear)*. PhD thesis, International Research Training Group Language Technology and Cognitive Systems, Saarland University.

Rieser, V. and Lemon, O. (2008a). Does this list contain what you were searching for? Learning adaptive dialogue strategies for interactive question answering. *Natural Language Engineering (special issue on Interactive Question Answering, to appear)*.

Rieser, V. and Lemon, O. (2008b). Learning effective multimodal dialogue strategies from Wizard-of-Oz data: Bootstrapping and evaluation. In *Proc. of the 21st International Conference on Computational Linguistics and 46th Annual Meeting of the Association for Computational Linguistics (ACL/HLT)*.

Steedman, M. and Petrick, R. (2007). Planning dialog actions. In *Proc. of the 8th SIGdial Workshop on Discourse and Dialogue*.

Young, S. (2000). Probabilistic methods in spoken dialogue systems. *Philosophical Trans Royal Society (Series A)*, 358(1769):1389–1402.

Zue, V. (2007). Organic interfaces. In *Proc. of the International Conference of Spoken Language Processing (Interspeech/ICSLP)*.

## Biographical Sketch

Since April 2008 the author is a Research Fellow at the School of Informatics, University of Edinburgh, where she carries out her post-doctoral research within the CLASSiC project (Computational Learning in Adaptive Systems for Spoken Conversation) funded by the EU FP7 Programme under grant agreement 216594 (`www.classic-project.org`). Prior to that she received a Ph.D. from Saarland University where she studied at the International Post-Graduate College for Language Technology and Cognitive Systems. Her thesis was supervised by Dr. Oliver Lemon (School of Informatics, Edinburgh) and Prof. Manfred Pinkal (Computational Linguistics, Saarbrücken). In 2004 she received an M.Sc. in "Language Engineering" from the School of Informatics, Edinburgh She also took an M.A. in Information Science and Linguistics at the University of Regensburg, Germany, in 2003.

# Robert J. Ross

Universitat Bremen
Enrique-Schmidt-Str. 5
28359 Bremen
Germany

robert.ross@informatik.uni-bremen.de
www.informatik.uni-bremen.de/~robertr/

## 1 Research Interests

Having started my research in the area of deliberative control for speech enabled agents [Ross et al., 2004a,b], my primary research interest is now focused within the Spoken Dialogue System (SDS) community – with emphasis on **Ontology & Semantics**, **Dialogue Structure & Control**, and **Dialogue System Integration**.

### 1.1 Ontology & Semantics

My main research interest has been the application of upper-ontologies to linguistic semantics and domain modelling for dialogue systems. I have pursued this through collaborative work on a Description Logic based Linguistic Ontology which forms the semantic interface to Categorial and Functional grammars. Part of that linguistic ontology has been the forthcoming Generalised Upper Model Version 3 [Bateman et al., 2006], while other parts have concerned inter-personal aspects (described below). I am also interested in the related topic of domain modelling of conceptual action and spatial representation for dialogue systems [Ross et al., 2006, Shi et al.].

### 1.2 Dialogue Structure & Control

Another active research interest of mine is dialogue structure and control modelling. My motivations for this research have been: (a) to understand the control mechanisms behind information state and agent-oriented approaches to dialogue management; and (b) to produce dialogue systems more suited to natural interaction through empirical investigations of the relationship between human-human and human-machine interaction at the level of speech acts and Generalised Dialogue Structures [Ross et al., 2005]. Following on from the ontology centric approach described above, I am currently formulating the interaction structures in terms of an Upper Interaction Ontology – an ontological module which forms a linguistic resource within a complete SDS.

### 1.3 Dialogue System Integration

I have also been interested in dialogue system integration for many years. While in early approaches I investigated a loose-coupled agent-centric integration strategy [Krieg-Brückner et al., 2004], in recent work I have pursued a more tightly coupled integration methodology. The principle result of that work is Corella, an open-source Java based dialogue management and integration library. That work, along with the other ideas outlined above, have been applied within a number of research prototypes including a speaking robotic wheelchair, and simulated conversational systems in the spatial domain.

## 2 Future of Spoken Dialog Research

I see the coming decade as a period of consolidation in practical dialogue research, but one in which a number of challenges and risks may yet present themselves for the dialogue systems community.

Facilitated by the direct incorporation of speech synthesis and recognition into commodity platforms and operating systems, Command & Control applications are likely to continue to grow in the applied system community. While positive from many perspectives, there is the potential for negative impact on the SDS community as a perception of *sufficient functionality having been reached* hampers further basic research funding. It may therefore be important for SDS researchers to pay measured attention to both practical system development in industry, as well as fundamental research in the lab.

As part of the proliferation of dialogue projects, consolidation in research systems may be possible as preferred stacks of language technologies develop. Young researchers can achieve much in moving towards this goal, but care should be taken to ensure that the results are not prescribed premature standardisation strategies. Instead, a natural consensus can be developed as common regularities in component and algorithm design are identified.

One major risk, but potential research possibility, may come in the area of speech recognition. Speech technology quality continues to be a bottleneck in SDS development leading both to perceptions of poor quality in system design, and considerable complications in SDS evaluation. It may be necessary for the upcoming SDS research community to devote much effort to the improvement of language technologies through the augmentation of language models with various forms of context, as well as the pursuit of parallelism and tight-coupling in the speech technology tool chain. These goals may not be achievable through ad-hoc efforts within general dialogue projects, but may instead have to be pursued explicitly as a dedicated research area.

Of course, opportunities for improvements in SDS research may come from any direction. Reasoning techniques researched within the Semantic Web community are likely to continue to be a fruitful source of basic components for SDS researchers. Similarly, research on user system design coming from the human-machine interaction community may provide empirical findings and design goals which should be built upon by dialogue systems researchers.

## 3 Suggestions for Discussion

**Coupling in Dialogue System Integration** Due to third-party and legacy software re-use, there has been a tendency to favour loose-coupled system integration facilitated by agent, middleware, or service architectures in SDS. While this approach has led to off-the-shelf integration strategies which allow rapid system development, an argument might be made that many limitations in current dialogue systems can only be solved through a far tighter coupling of the components which compose an SDS. Issues for discussion in this topic might include an analysis of trends in this area, and an evaluation of how tighter coupling might alleviate current difficulties in wide-coverage speech recognition.

**Safety & Reliability in Dialogue Systems** While music selection and GPS navigation enquiries are common deployments of dialogue systems to safety critical environments such as the automotive industry, it is likely that SDS will be further integrated into key navigation and control systems in both that automotive and service robotics industries over time. These practical applications present a number of interesting questions for dialogue system researchers as we move beyond research lab prototypes: What practical or regulatory challenges do safety critical applications present to the SDS community? Should conscious compromises in functionality be made to guarantee predictability in operation? And are dialogue systems at all mature enough to be deployed to critical system applications?

**Benchmarking SDS Research** With large funding bodies such as the EU Framework Programme and Germany's DFG increasingly requesting benchmarking of research and development beyond publication statistics, the SDS community may need to propose clear benchmarking methodologies to satisfy these funding agencies. Issues such as inter-project and complete system evaluation are of course relevant here, but so too are intra-project issues such as resource development and code reuse. Questions for discussion would include: Are explicit benchmarking requests already making their way into SDS funding strategies? What can we learn from outside the SDS community? And what coherent suggestions for benchmarking can we make in a research field which is notoriously difficult to evaluate?

## References

J. Bateman, S. Farrar, J. Hois, and R. Ross. The Generalized Upper Model 3.0: Documentation. SFB/TR8 internal report, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany, 2006.

Bernd Krieg-Brückner, Hui Shi, and Robert Ross. A safe and robust approach to shared-control via dialogue. *Journal of Software*, 15(12):1764 –1775, 2004.

Robert Ross, Rem Collier, and G.M.P. O'Hare. Demonstrating social error recovery with agentfactory. In *Proceeedings of The Third International Joint Conference on Autonomous Agents and Multi Agent Systems*, 2004a.

Robert Ross, Rem Collier, and G.M.P. O'Hare. AF-APL Bridging Princples & Practices in Agent Oriented Languages. In *PROMAS 2004*, volume 3346 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2004b.

Robert J. Ross, John Bateman, and Hui Shi. Using Generalised Dialogue Models to Constrain Information State Based Dialogue Systems. In *the Symposium on Dialogue Modelling and Generation, 2005*, Amsterdam, The Neterhands., 2005.

Robert J. Ross, Christian Mandel, John Bateman, Shi Hui, and Udo Frese. Towards stratified spatial modeling for communication and navigation. In *IROS Workshop From Sensors to Human Spatial Concepts 06*, Beijing, China, 2006.

Hui Shi, Christian Mandel, and Robert J. Ross. Interpreting route instructions as qualitative spatial actions. In *Spatial Cognition V*, Lecture Notes in Artificial Intelligence. Springer Verlag; 14197 Berlin; http://www.springeronline.com.

## Biographical Sketch

Robert J. Ross is a research assistant within the Spatial Cognition Collaborative Research Center (SFB/TRR) Bremen, Germany. Robert completed an MSc in 2003 on Robotic System Integration in the Computer Science Department of University College Dublin, Ireland. Since moving to the Collaborative Research Center he has worked in the area of dialogue systems for robotic agents within the projects I3-[SharC] and I5-[DiaSpace]. He is currently writing up his PhD thesis on spatial language contextualisation within practical dialogue systems.

**Mihai Rotaru**

University of Pittsburgh
5420 Sennott Hall
210 S. Bouquet
Pittsburgh, PA 15260, USA

mrotaru@cs.pitt.edu
www.cs.pitt.edu/~mrotaru

## 1 Research Interests

My research focuses on **empirical approaches** to spoken dialogue system **analysis** and **design**. I am primarily interested in the challenges and the opportunities that arise from **adding speech interfaces to complex domains** (e.g. tutoring, troubleshooting).

The testbed of my research is the **ITSPOKE** system developed at University of Pittsburgh within Diane Litman's research group. During my Ph.D. studies I was involved in several research projects using this system. Among others, I looked at emotion prediction using sub-turn features (Nicholas et al., 2008; Rotaru and Litman, 2005) and at interactions between dialogue phenomena (Rotaru and Litman, 2006a, 2006b). My dissertation work investigates and validates the utility of discourse structure for spoken dialogue systems in complex domains and a brief description of its current status is available below.

### 1.1 Applications of Discourse Structure

Dialogues (human-human or human-computer) have an inherent structure called the discourse structure. To make an analogy, discourse structure is for dialogues what an outline is for a textbook. However, due to the relatively simple structure of dialogues in previous spoken dialogue systems, discourse structure has seen limited applications. In contrast, dialogues in complex domains like tutoring or troubleshooting exhibit a richer discourse structure which enables new applications of this concept.

Two types of applications are being pursued: on the system side and on the user side. This classification reflects the direct beneficiary of the applications: the system designer or the user, respectively.

On the system side, my work investigates the applications of discourse structure for performance analysis and characterization of dialogue phenomena. The task of performance analysis is to discover factors that relate to or impact the system performance. Two intuitions have guided our use of discourse structure. First, phenomena related to performance (e.g. speech errors, user emotions, etc) are not uniformly important across the dialogue but have more weight at specific places in the dialogue. We use transitions in the discourse structure to define the notion of "places in the dialogue". Second, "good" dialogues have a discourse structure different from "bad" dialogues. An empirical study (Rotaru and Litman, 2006c) validates these intuitions and provides important insights about factors that relate to system performance (i.e. behavior after certain discourse structure transitions is associated with increased system performance). A modification of our system was informed by these insights and a current study investigates its performance improvements.

For the characterization of dialogue phenomena, we hypothesize that these phenomena are not uniformly distributed across the dialogue but are more frequent at specific places in the dialogue. Again, we use discourse structure transitions to define the notion of "places in the dialogue". Empirical studies confirm this hypothesis for speech recognition problems (Rotaru and Litman, 2006b) and for user affect (Forbes-Riley et al., 2007). We find that certain transitions are associated with an increase in speech recognition problems or user affect. From the dialogue designer perspective, these results suggest that particular attention should be paid at specific locations in the discourse structure. In addition, the observed interactions suggest that discourse structure can be an informative feature for predicting speech recognition problems or user affect.

On the user side, we hypothesize that it is easier for users to follow the conversation with a system if a graphical representation of the discourse structure is present. This representation is called the Navigation Map. We conducted a user study focused on the user's perception of the system with and without the Navigation Map (Rotaru and Litman, 2007). An analysis of users' ratings indicates that they prefer the Navigation Map-enabled version on various dimensions. The Navigation Map presence allows users to better identify and follow the tutoring plan and to better integrate the instruction. It was also easier for users to concentrate and to learn from the system if the Navigation Map was present. We are currently conducting another user study that investigates the objective utility of the Navigation Map.

## 2 Future of Spoken Dialog Research

Dialogue systems for call-centers and in-car entertainment/navigation have matured over the past few years and there are many examples of successful commercialization. In my view, the next step for spoken dialogue systems is to go back to their roots and become once again "personal". By personal, I mean systems that cater to the informational, organizational and entertainment needs of their users. In-car systems already fit in this category and there are a number of startups that are exploring mobile phones as the vehicle for such systems (e.g. SpinVox). The CALO project (caloproject.sri.com) is pursing a similar idea.

Another novel application for dialogue systems is data processing. While it has become increasingly easy for people to store and access information, I believe that the tools for processing information are inadequate. The problem with current tools is that they require a long time to train and that users are always struggling to match their processing needs to the available functionalities. A better approach will be to allow users to describe in (spoken) natural language what they want to do and to disambiguate if needed. I envision a dialogue system that incrementally acquires user's description of the data, verifies consistency of the description with the data, provides an arsenal of predefined data processing operations and allows users to define new operations. This domain raises many interesting research questions related to the acquisition of vocabulary/operations from user, natural language understanding, dialogue management, natural language generation and constraint checking.

To summarize, I believe we are seeing for spoken dialogue systems the same transition we saw for computer systems 25 years ago: moving from the enterprise paradigm to the personal paradigm.

## 3 Suggestions for discussion

- What data analysis methods should we use to identifying system's weak points from collected corpora? Can these analyses be performed with limited or no human feedback?
- There is a gap between the technologies used in research and industry. What is the impact of this gap for young researchers applying for jobs in industry and how to address it?
- What are the tasks where *personal* spoken dialogue systems will have the most impact and lead to the largest adoption?

## References

K. Forbes-Riley, M. Rotaru, D. Litman and J. Tetreault. 2007. *Exploring Affect-Context Dependencies for Adaptive System Development.* In Proc. of Human Language Technology / North American Chapter of the Association for Computational Linguistics Conference (HLT/NAACL).

G. Nicholas, M. Rotaru and D. Litman. 2008. An Investigation of Using Word-level Features for Emotion Prediction. *Speech Communication, submitted.*

M. Rotaru and D. Litman. 2005. *Using Word-level Pitch Features to Better Predict Student Emotions during Spoken Tutoring Dialogues.* In Proc. of European Conference on Speech Communication and Technology (Interspeech-2005/Eurospeech).

M. Rotaru and D. Litman. 2006a. *Dependencies between Student State and Speech Recognition Problems in Spoken Tutoring Dialogues.* In Proc. of ACL.

M. Rotaru and D. Litman. 2006b. *Discourse Structure and Speech Recognition Problems.* In Proc. of Interspeech.

M. Rotaru and D. Litman. 2006c. *Exploiting Discourse Structure for Spoken Dialogue Performance Analysis.* In Proc. of EMNLP.

M. Rotaru and D. Litman. 2007. *The Utility of a Graphical Representation of Discourse Structure in Spoken Dialogue Systems.* In Proc. of ACL.

## Biographical Sketch

Mihai Rotaru is a Ph.D. candidate in the Department of Computer Science, University of Pittsburgh, USA. He works under the supervision of Dr. Diane J. Litman. He was born in Romania and received his B.Sc. and M.Sc. from West University, Timisoara, Romania. His other interests include good movies, reading, traveling, cool gadgets, friends and (last but not the least) good beer. These interests combine sometimes with his academic interests at conferences (especially during evenings).

**Umar Syed**

Princeton University
Department of Computer Science
35 Olden St.
Princeton, NJ 08540, USA

usyed@cs.princeton.edu
www.cs.princeton.edu/~usyed

## 1 Research Interests

My research in spoken dialog systems has focused on **learning user models**. More generally, I am interested in **machine learning**, especially **reinforcement learning**. Recently, I have begun some work in **game theory**.

### 1.1 Learning User Models From Unlabeled Dialog Data

In the machine learning approach to dialog management, the problem of building a good dialog manager is reduced to the problem of finding an optimal policy in a (Partially Observable) Markov Decision Process (Levin et al., 2000; Williams and Young, 2007). Since algorithms for policy learning in (PO)MDPs often require a lengthy and costly exploration phase, running these algorithms on real users is impractical. For this reason, it is desirable to build realistic models of user behavior with which experimentation can be freely done.

User models are typically estimated from transcribed corpora of human-computer dialogs. Of course, because the transcriptions are done manually, these corpora are usually small and sparse. In (Syed and Williams, 2008), we describe a method for learning user models that operates on dialogs that have been automatically transcribed by an ASR engine. Since the ASR process is error-prone, we cannot assume that these transcripts will accurately reflect users' true actions and internal states. To handle the uncertainty, we employ an EM algorithm that treats this information as unobserved data. The EM algorithm infers the presence of transcription errors by using an ASR confusion model that is estimated from a very small number of manually transcribed dialogs. Our experiments indicate that our method learns user models that are very similar to those learned by simpler methods that use a much larger set of manually transcribed dialogs.

### 1.2 Speeding Up the Learning of User Models with a "Value-Based" Prior

The approach described in the previous section learns a user model by finding a maximum likelihood (ML) estimate from data. Incorporating prior knowledge about user behavior can help improve this estimate. If we can restrict the space of user models to just those that agree with our pre-existing beliefs about how users tend to interact with a spoken dialog system, then hopefully less data will be needed for learning.

Detailed prior knowledge about user behavior is usually hard to obtain; this is, after all, the very problem we are trying to solve. But for most spoken dialog systems, it is fair to assume that users are acting in goal-directed manner (see e.g. (Scheffler and Young, 2001)). They may be booking a flight, retrieving a directory listing, or trying to accomplish some other task. In (Syed and Schapire, 2007) we assume that, for an MDP that models the dialog generating process, we can design a reward function that is consistent with how users behave. For example, we can assign higher rewards to dialog states that are nearer to fulfilling the user's objective. Then we identify the space of possible user models with the set of high-value policies in this MDP. In this way, we can leverage our prior knowledge that users are trying to complete their conversations as soon as possible, without needing to specify exactly how they are trying to accomplish that goal.

Instead of finding the ML estimate, our learning algorithm seeks a maximum *a posteriori* (MAP) estimate for the user model, in which the likelihood is combined with our "value-based" prior. Our algorithm provably converges to a stationary point of this posterior.

### 1.3 Future Work: Directly Optimizing User Model Evalutation Metrics

A major challenge in the design and analysis of user models is determining whether one user model is more "realistic" than another. There is unfortunately no widely-accepted evaluation metric for measuring this property. See (Schatzmann et al., 2005; Williams, 2007) for discussion and proposed solutions. In addition to measuring realism, a user model evaluation metric ought to be "optimizable": it should be possible to design efficient algorithms that learn user models that maximize the value of the metric. Indeed, this is how any proposed metric will ultimately be used. However, relatively little attention has been paid to this issue.

To motivate the issue further, let us describe how the use of a suitable evaluation metric can lead to a considerable savings in the amount of data need to learn a user model. For a particular spoken dialog system (SDS), let $\mathcal{D}$ be the distribution on dialogs induced be real users interacting with this SDS, and let $\mathcal{D}'$ be the distribution on dialogs induced by a user model interacting with the same SDS. Suppose that if

$$E_{d\sim\mathcal{D}}[f(d)] \approx E_{d'\sim\mathcal{D}'}[f(d')], \qquad (1)$$

then the user model is deemed to be realistic, where $f(\cdot) \in \mathbb{R}$ is some property of dialogs (length, number of confusions, etc.). Putting aside the difficulty of designing an appropriate $f(\cdot)$, it can be shown that if each real dialog $d$ is drawn independently from $\mathcal{D}$ (a reasonable assumption for most dialog applications), then $E_{d\sim\mathcal{D}}[f(d)]$ can be estimated very accurately using a number of real dialogs that is completely independent of the complexity of user behavior or of the SDS. So it seems possible, at least in principle, to design a learning algorithm that requries very little data and produces a user model whose distribution $\mathcal{D}'$ satisfies (1). I believe the techniques described in (Syed and Schapire, 2008; Syed et al., 2008) can be applied here.

## 2 Future of Spoken Dialog Research

Generally speaking, I believe that machine learning techniques will further migrate into spoken dialog research. The spoken dialog problem domain shares many properties with other domains in which machine learning has been successful: expensive and noisy data, notions of correctness that are easier to demonstrate than describe, etc.

## 3 Suggestions for Discussion

- As I asked in Section 1.3, is it possible to define evaluation metrics for user models that not only measure realism, but can also be directly optimized by efficient learning algorithms?

- The MDP framework has become a widely-applied approach to learning dialog managers. But the earliest and probably most successful machine learning paradigm is supervised learning. Moreover, demonstrations from good dialog managers are plentiful, though very noisy: human-human dialog transcripts. Are there supervised learning techniques that are suitable for learning from this data, e.g. structured prediction?

- Dialog managers and user behavior are usually modeled as functions of the dialog state, a (relatively) low-dimensional object which is supposed to summarize all the relevant details of the dialog. Current methods for defining the dialog state are fairly ad-hoc. Are there more principled methods, particularly machine learning methods, that can be applied here?

## References

E Levin, R Pieraccini, and W Eckert. 2000. A stochastic model of human-machine interaction for learning dialogue strategies. *IEEE Trans on Speech and Audio Processing*, 8(1):11–23.

J Schatzmann, K Georgila, and SJ Young. 2005. Quantitative evaluation of user simulation techniques for spoken dialogue systems. In *Proc SIGdial Workshop on Discourse and Dialogue, Lisbon, Portugal*, pages 178–181.

K Scheffler and SJ Young. 2001. Corpus-based dialogue simulation for automatic strategy learning and evaluation. In *Proc North American Association for Computational Linguistics (NAACL) Workshop on Adaptation in Dialogue Systems, Pittsburgh, USA*, pages 12–19.

U Syed and RE Schapire. 2007. Imitation learning with a value-based prior. In *Proc Conf on Uncertainty in Artificial Intelligence (UAI), Vancouver, BC, Canada.*

U Syed and RE Schapire. 2008. A game-theoretic approach to apprenticeship learning. In *Proc Advances in Neural Information Processing Systems (NIPS), Vancouver, BC, Canada*, pages 1635–1642.

U Syed and JD Williams. 2008. Using automatically transcribed dialogs to learn user models in a spoken dialog system. In *Proc Association for Computational Linguistics (ACL), Columbus, USA.*

U Syed, M Bowling, and RE Schapire. 2008. Apprenticeship learning using linear programming. In *Proc Intl Conf on Machine Learning (ICML), Helsinki, Finland.*

JD Williams and SJ Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

JD Williams. 2007. A method for evaluating and comparing user simulations: The Cramer-von Mises divergence. In *Proc IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan.*

## Biographical Sketch

Umar Syed is a fourth-year Ph.D. student in the Department of Computer Science at Princeton University. His advisor is Rob Schapire. Last summer, he was a research intern at AT&T Labs in Florham Park, NJ, where he was mentored by Jason Williams. He has a M. Eng. in Computer Science from Cornell University and a B.Sc. in Computer Engineering from the University of Florida.

# Jason D. Williams

AT&T Labs – Research
Shannon Research Lab
180 Park Ave
Florham Park, NJ, USA

jdw@research.att.com
www.research.att.com/info/jdw

## 1 Research Interests

My main research interests are building dialog systems that are resilient to speech recognition errors and which can cope with variability in user behavior. More specifically, I am interested in **robust dialog modelling**, or accurately tracking the state of a human-computer dialog, and to support this I am interested in **user simulation** and **ASR modelling**. Since the state of the dialog can never be know for sure, choosing actions is challenging. I believe that **optimization-assisted dialog management**—a combination of human design expertise and automatic optimization—is essential. This draws on both conventional design techniques and **planning under uncertainty**, which also interests me in general.

### 1.1 Robust dialog modelling

Conventional systems track a single hypothesis for the dialog state. This formulation struggles to make full use of the N-Best list and discards useful cross-turn information. By contrast, a distribution over a multiple dialog state hypotheses adds inherent robustness, because even if an error is introduced into one dialog hypothesis, it can later be discarded in favor of other, uncontaminated dialog hypotheses.

I have developed a method to represent multiple dialog hypotheses in the form of a Bayesian network, called the SDS-POMDP model (Williams and Young, 2007a). A hidden state is decomposed into three elements: the user's goal, the user's action, and the dialog history. Models of how users behave and how speech recognition errors are introduced can be estimated from data, and this allows a dialog system to track a distribution over all possible user goals as the dialog progresses. I initially developed this model for slot-filling dialogs and have also applied it to the troubleshooting domain (Williams, 2007a).

Updating this distribution must be done in real-time, and as the number of possible hidden states grows, this becomes impossible to do exactly. To address this I have also explored ways to update this distribution efficiently and approximately, for example using particle filters or a dialog beam (Young et al., 2006; Williams, 2007c).

### 1.2 Optimization-assisted dialog management

In conventional systems, the dialog plan is designed by a developer. While this helps incorporate domain knowledge, a developer can only consider a relatively small number of dialog situations. By contrast, applying automatic optimization to the dialog manager allows many more conversational situations to be considered than is feasible by hand, and this creates much more detailed and effective dialog plans (Williams and Young, 2007a). Automatic optimization in this context is usually intractable without some form of compression, and this has prompted work to compress dialog management problems into more compact representations (Williams and Young, 2007b; Williams and Young, 2006).

These optimizations are done off-line, with a user model. For the optimizations to improve performance with real users, it is crucial that the user model is a faithful representation of real users. I have worked to understand how users behave in real systems (Williams and Witt, 2004; Williams and Young, 2004), and also developed a quality measure for user simulations, based on the Cramér-von Mises divergence (Williams, 2007b).

### 1.3 Future work

Currently, I am working to extend these approaches in several ways. First, I'm looking at how to train user models more accurately and with less effort. Second, I am pursuing more sophisticated methods for representing multiple hypotheses for the dialog state. Finally, I am investigating how robust dialog modelling and optimization-assisted dialog management can be made compatible with conventional practices, and accessible to the practitioners in industry.

## 2 Future of Spoken Dialog Research

First, mobile computing is becoming the norm. Always-on, always connected devices are nearly general-purpose computers, yet they lack a full-size computer's usability. Spoken dialog systems could play a vital role in improving the usability of these devices.

Second, people in the developed world will be aging in record numbers and the costs of assisting and caring for the elderly will rise. Many care tasks could be provided by a machine or robot, and this group of users might find a conversational interface more helpful than a GUI.

Finally, energy costs are increasing in most parts of the world. Perhaps interactive and aware systems in the home could help reduce consumption – for example, an in-home dialog system could ask permission to turn off lights in other parts of the house, or use heating more efficiently by asking how long someone intends to be away from home.

To address these types of applications, I believe our field needs to look carefully at a few core research issues, especially robustness to speech recognition errors. Speech recognition and understanding errors will not disappear and robustness at the dialog level must be fundamental to dialog management just as a beam search is fundamental to ASR decoding.

## 3 Suggestions for Discussion

- **Speech interfaces on mobile devices**: The convergence of mobile data networks with powerful pocket-size devices appear to be a superb opportunity for speech interfaces. How are people likely to use mobile services, what role can spoken dialog systems play, and what are the enabling innovations needed?

- **An X-Prize for spoken dialog systems**: Currently it is difficult to compare different approaches to dialog systems because different research groups work on different problems. Challenge problems have historically been excellent community motivators. Is there a role for an X-Prize for spoken dialog systems? How could such a competition be structured?

- **Wider adoption of statistical techniques**: Statistical techniques such as tracking a distribution over multiple dialog states or using reinforcement learning have shown good promise in research settings. However, these methods have not (yet!) been adopted in industry. How can these methods be made more accessible and incorporated into conventional dialog design practices?

## References

JD Williams and SM Witt. 2004. A comparison of dialog strategies for call routing (invited article). *International Journal of Speech Technology*, 7(1):9–24.

JD Williams and SJ Young. 2004. Characterizing task-oriented dialog using a simulated ASR channel. In *Proc Intl Conf on Spoken Language Processing (ICSLP), Jeju, Korea*, pages 185–188.

JD Williams and SJ Young. 2006. Scaling POMDPs for dialog management with composite summary point-based value iteration (CSPBVI). In *Proc American Association for Artificial Intelligence (AAAI) Workshop on Statistical and Empirical Approaches for Spoken Dialogue Systems*.

JD Williams and SJ Young. 2007a. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech and Language*, 21(2):393–422.

JD Williams and SJ Young. 2007b. Scaling POMDPs for spoken dialog management. *IEEE Trans. on Audio, Speech, and Language Processing*, 15(7):2116–2129.

JD Williams. 2007a. Applying POMDPs to dialog systems in the troubleshooting domain. In *NAACL-HLT Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies, Rochester, New York, USA*, pages 1–8.

JD Williams. 2007b. A method for evaluating and comparing user simulations: The Cramer-von Mises divergence. In *Proc IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan*.

JD Williams. 2007c. Using particle filters to track dialogue state. In *Proc IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Kyoto, Japan*.

SJ Young, JD Williams, J Schatzmann, MN Stuttle, and K Weilhammer. 2006. The hidden information state approach to dialogue management. Technical Report CUED/F-INFENG/TR.544, Cambridge University Engineering Department.

## Biographical Sketch

Jason D. Williams is a Principal Member of Technical Staff with AT&T Labs – Research in Florham Park, New Jersey, USA. He received a BSE in Electrical Engineering from Princeton University in 1998, and at Cambridge University he received an M Phil in Computer Speech and Language Processing in 1999 under a Churchill Scholarship and a Ph D in Information Engineering in 2006 under a Gates Scholarship. He has previously held positions at Tellme Networks (now Microsoft), Edify Corporation (now Intervoice), and McKinsey & Company's Business Technology Office. Commercially deployed dialog systems he has built for companies such as Sony, BMW, and AT&T currently receive millions of calls per year.

**Andi Winterboer**

University of Edinburgh
School of Informatics, ICCS
Edinburgh, United Kingdom

A.Winterboer@ed.ac.uk

# 1 Research Interests

My general research interests lie in the broad areas of **human-computer interaction**, **cognitive science** and **natural language processing**. Currently, I am examining different **information presentation strategies** of spoken dialogue systems (SDS). In particular, I aim to assess the effect of information strategies on user perception, task success, cognitive load, and user recall of information.

## 1.1 Past work

Previously, I worked on developing and evaluating speech interfaces to enable intuitive robot control (Tenbrink and Winterboer, 2005). The main focus of my work has been to explore how humans interact with robots in order to communicate spatially relevant information. Therefore, I iteratively developed a speech interface for an AIBO robot (Winterboer et al., in press).

## 1.2 Current and future work

Although a lot of research has been done on the information gathering phase of spoken dialogue systems, there is relatively little done regarding information presentation. Yet, analyzing the Communicator corpus consisting of approximately 2000 dialogues with nine different spoken dialogue systems it was found that 69% of the dialogue when measured in time, and 91% when measured in words, is due to the system producing utterances (Moore, 2006). Therefore, it is crucial that we gain understanding of how best to design information presentation strategies for spoken dialogue systems. In addition, intelligent algorithms for effective information presentation have to be devised taking into account that many SDS are intended for use in situations where the user is performing another task simultaneously, e.g., riding on a train, walking, or driving a car. Especially in the context of in-car information systems safety is of paramount importance and distraction effects must be minimized in order to guarantee that driving behaviour is not adversely affected. Therefore, an assessment of the cognitive load imposed by different information presentation strategies is an important factor in the design and development of in-car voice services. To explore this question, we performed two experiments comparing two previously proposed approaches to information presentation, focusing on their effect on driving-related cognitive load. The summarize and refine approach (SR) to information presentation, developed by (Polifroni et al., 2003), groups a large number of options into a small number of clusters that share attributes. Then, the system summarizes the clusters based on their attributes and suggests additional constraints to the user. In contrast, the user-model based summarize and refine approach (UMSR) (Demberg and Moore, 2006) employs a user model to reduce dialogue duration by considering only options that are relevant to the user. When the number of relevant items exceeds a manageable number, the UMSR approach builds a cluster-based tree structure which orders the options for stepwise refinement based on the ranking of attributes in the user model. The effectiveness of the tree structure, which directs the dialogue flow, is enhanced by taking the user's preferences into account. In order to provide the user with a better overview of the option space, tradeoffs between alternative options are presented explicitly.

We conducted two Wizard-of-Oz experiments comparing these approaches to information presentation in situations of low vs. high workload with a simulated SDS (Hu et al., 2007), (Winterboer et al., 2007). In these experiments, participants interacted with a spoken dialogue system in the flight booking domain. Thus, we were able to assess the impact of the different approaches on effectiveness criteria such as task duration and completion. We found that the UMSR approach enables more efficient information retrieval in comparison with the SR approach, and that presenting information with UMSR did not negatively affect driving performance. However, in contrast to results of previous studies showing significant preferences for UMSR when participants were reading or overhearing dialogues, no differences between user satisfaction ratings of the two presentation methods were observed in the dual task studies. Thus, in order to find out whether the lack of differences between the user satisfaction ratings was caused by the fact that participants were actually conversing with a SDS, or whether the reason was the demanding secondary task, we carried out an additional experiment in which participants only interacted with the simulated SDS (Winterboer and Moore, 2007). The results of this experiment seem to suggest that the secondary task did affect user ratings. Possibly, participants in conditions of high cognitive load are so concerned with completing the dual tasks that they are less aware of differences in wording or in the order in which options and their attributes are presented. Not only did the UMSR approach in this experiment again outperform SR in terms of task success and dialogue duration, and enabled more effective information retrieval, we also found user ratings to demonstrate a consistent trend favoring recommendations based on UMSR.

Typically, spoken dialogue systems present information about restaurants, flights, or products using relatively simple templates for natural language realization. Recently, however, a number of approaches to information presentation were introduced using discourse cues (e.g., but, however, moreover, only, just etc.) in order to highlight specific properties of and relations between the presented items (Demberg and Moore, 2006), (Winterboer and Moore, 2007). We performed a within-participants reading experiment (Winterboer and Moore, 2008) comparing item recall for material presented with and without discourse cues. Overall, we found a consistent trend indicating that items in messages containing discourse cues could be recalled more easily. This finding shows that highlighting similarities and contrasts between different options can indeed facilitate the recall of information. In future work, we aim to examine whether these findings can be replicated with speech rather than text.

## 2 Future of Spoken Dialog Research

In the near future, mobile devices will become even smaller and more powerful. At the same time, traditional input methods, mainly clicking tiny buttons, lack the convenience users expect from todays mobile devices. Speech interfaces could certainly facilitate the whole interaction process by additionally taking into account the users' preferences, emotional and cognitive state. Moreover, SDS could play an increasing role in service robotics and video games.

## 3 Suggestions for Discussion

- Traditional user studies are useful, but expensive in costs and time. Are there other meaningful ways to receive input and feedback from users? Especially if we only want to test one SDS component (e.g., NLG module)?
- I second the idea of a dialogue challenge because such competitions were successful in other areas of NLP. The question is: How to design it and where to start?
- Nowadays, a lot of research focuses on correctly recognizing user utterances. However, enabling users to easily comprehend system utterances is similarly important. What, from your experience, are the factors that have the most negative impact on the comprehension of system utterances (e.g., sentence complexity, turn length, voice type, . . . )?

## References

V. Demberg and J. D. Moore. 2006. *Information Presentation in Spoken Dialogue Systems*. Proc. of the 11th Conference of the Europ. Chap. of the ACL (EACL '06). Trento, Italy.

J. Hu, A. Winterboer, C. Nass, J. D. Moore, and R. Illowsky. 2007. *Context & usability testing: User-modeled information presentation in easy and difficult driving conditions*. Proc. of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07). San Jose, California.

J. D. Moore. 2006. *Natural language generation for information presentation*. Talk at the Spoken Language Technology Workshop (SLT '06). Aruba.

J. Polifroni, G. Chung, and S. Seneff. 2003. *Towards automatic generation of mixed-initiative dialogue systems from web content*. In Proc. of Eurospeech '03. Geneva, Switzerland.

T. Tenbrink and A. Winterboer. 2005. *Spatial Directionals for Robot Navigation*. Thematic Session on Motion Encoding, 21st Scandinavian Conf. of Ling. Trondheim, Norway.

A. Winterboer, J. Hu, J. D. Moore, and C. Nass. 2007. *The Influence of User Tailoring and Cognitive Load on User Performance in Spoken Dialogue Systems*. Proceedings of Interspeech - ICSLP '07. Antwerp, Belgium.

A. Winterboer and J. D. Moore. 2007. *Evaluating Information Presentation Strategies for Spoken Recommendations*. In Proc. of the ACM Conf. on Recommender Systems '07. Minneapolis, MN.

A. Winterboer, T. Tenbrink, and R. Moratz In press. *Spatial Directionals for Robot Navigation.*. In E. van der Zee and M. Vulchanova (eds.) Motion Encoding in Language and Space. Oxford University Press.

A. Winterboer and J. D. Moore. 2008. *Do discourse cues facilitate recall in information presentation messages?* Submitted.

## Biographical Sketch

Andi Winterboer is a PhD student at ICCS at the University of Edinburgh. Prior to his PhD studies, Andi received a diploma in Informatics from the University of Bremen, Germany, in 2004. He is funded by the Edinburgh-Stanford Link and works under the supervision of Johanna Moore and Fernanda Ferreira.

# Sabrina Wilske

Department of Computational Linguistics
Saarland University
Saarbrücken, D-66041

sw@coli.uni-sb.de

## 1 Research Interests

My research interest lies in two areas within the broad field of dialog systems, one area is dialog in **human-robot interaction**, the other is dialog for **language learning**. In the first area we try to enable robots to communicate with humans using speech, where the embodiment of the robot and the situatedness of the interaction give rise to a range of research issues that non-robot dialog systems don't have to address. In the second area we deal with dialog systems that complement or substitute human teachers or conversation partners for the goal of learning a foreign language. On the one hand these systems should be able to talk about any domain (which requires general and widespread knowledge), on the other hand they also need explicit and specific knowledge about the language in order to correct the learners' errors and direct their attention to specific, syllabus-driven issues.

### 1.1 Dialog systems for human robot interaction

My past work focused on dialog between robots and humans. I worked in the CoSy-project (Cognitive Systems for Cognitive Assistants) where we developed embodied conversational agents, i.e. talking robots. The robots should not only be able to talk, but also be able to learn how to relate language, action and situated reality. For my master's thesis (Wilske, 2006), I tried to enable robots to learn how to communicate with humans, more specifically I made a robot learn (1) to understand indirect speech acts and (2) how to engage humans to help it in pursuing its goals. Both tasks require linguistic knowledge and an appreciation of the reactions of the communication partners.

**Understanding indirect speech acts** People don't always literally mean what they say: They say $A$, but indirectly really mean $B$ – and $B$ is what you ought to do. We say that such utterances express *indirect speech acts* (ISAs) (Searle, 1969). Understanding the motives behind what someone says is particularly crucial in service-oriented human-robot interaction (HRI), where a robot often needs to act on the basis of what the human tells it. We developed an approach to interpret request-ISAs in human robot interaction. The robot interprets utterances based on their linguistic meaning and classification in the context of the current situation. If an utterance is initially ambiguous the robot asks for clarification and adapts its interpretation strategy. The key features of our approach are situation-dependent flexibility and adaptivity. See (Wilske and Kruijff, 2006) for more details.

**Requesting help** A single robot often has limited capabilities to act upon its environment, which demands an ability to collaborate with other agents. If the robot is to collaborate with human partners, it should have means to coordinate such cooperation using natural language. We enabled the robot to request help in case it couldn't pursue its goal on its own. The robot had at its disposal a small set of possible request utterances that differed in politeness realized by varying sentence mood. According to the reactions of the helper-to-be the robot learns to use the most successful utterance.

### 1.2 Dialogs for computer-assisted language learning

I started to work on my PhD in 2007. I am focusing now on modeling and managing dialogs for computer assisted language learning (CALL). Dialog systems for CALL are often motivated by the communicative approach to second language acquisition, which claims that learners profit from solving non-language problems using the developing second language (Douglas, 1995).

Tutorial dialog systems within this paradigm usually build tasks and domain descriptions along with associated microworlds or a role-play setting. These systems are normally built manually, sometimes they come with a drag-and-drop style authoring tool (Holland et al., 1995). Usually, creating a dialog model and the corresponding systems or adapting an existing model to another domain or language requires some amount of designing effort and specialized knowledge of the dialog modeling formalism and the knowledge representation language which is used for encoding the domain-specific knowledge.

One goal of my thesis is to develop mechanisms and a tool that allow non-experts in dialog systems to develop dialog models. The tool should allow language teachers or authors of language learning syllabi, or even the learner themselves to create and customize dialog models and use them for practicing communicative skills.

On our way to that goal, we are currently developing on an architecture for authoring and executing language learning dialogs. The architecture (i) allows learners to exercise dialogs in various scenarios and their different variants within one scenario, (ii) allows teachers without technical background to author the dialogs and the necessary linguistic resources, (iii) is modular in a way that facilitates division of labor between content authors with and without technical background. The key of our approach is to separate domain knowledge from the dialog structure and the language model and thereby facilitating a division of the authoring task.

Considering that dialogs for communicative skill practice can be about any desired topic and thus should be very broad and general regarding the knowledge they cover, we need methods that allow us to develop dialog models from different kinds of knowledge sources. One part of our work is to investigate what kinds of knowledge representation are necessary and appropriate for different kinds of dialog scenarios. The next step is then to come up with easy and intuitive ways to author that knowledge, possibly with a general knowledge base as support.

## 2 Future of Spoken Dialog Research

In my opinion a major problem in current work on spoken dialog systems is the missing robustness of automatic speech recognition. The amount of speech recognition errors is a big hindrance to the further propagation of dialog systems. Thus I think progress in automatic speech recognition is pivotal for SDS.

Another problem is the lack of tools for building dialog systems. At the moment, dialog systems are built for one specific application, and every time you need a new application you have to build a new dialog system from scratch. Although there has been important work on tools that allow you to build systems (DIPPER[1], TRINDIKIT[2], ARIADNE[3]) these are still relatively limited. In my opinion further research and engineering in this field is needed.

A third issue that is worthwhile to consider is the knowledge acquisition problem: How can we automatically design dialog systems using existing sources of knowledge, like manuals, FAQ lists, or other semi-structured information available on web pages.

## 3 Suggestions for Discussion

- Scalability, re-usability and transferability of dialog systems: How can we go beyond building one specific system that can only be used for a specific task and develop methods and tools that enable us to easily build a range of different dialog systems with different domains and task structures.
- Social competence for dialog systems: How can we endow dialog systems with social competence, make them polite, maybe even empathetic?
- Ethical responsibility of researchers: This is a topic that is not specific to research in dialog systems but to scientific research in general. Results of scientific research don't always serve the betterment of humanity, but sometimes cause misery and pain. Developing dialog systems might not have the same impact as inventing gun powder or the atomic bomb, but still I think that we should be aware of the consequences of our work and how we might contribute voluntarily or involuntarily to undesirable states of the world, e.g. repression of people.

## References

Sarah A. Douglas, 1995. *Intelligent Language Tutors - Theory Shaping Technology*, chapter LingWorlds: An intelligent object-oriented environment for sec- ond language tutoring. Lawrence Erlbaum Associates.

V. Melissa Holland, Jonathan D. Kaplan, and Michelle R. Sams, editors. 1995. *Intelligent Language Tutors: Theory Shaping Technology*. Lawrence Erlbaum Associates.

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge, United Kingdom.

Sabrina Wilske and Geert-Jan Kruijff. 2006. Service robots dealing with indirect speech acts. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 4698–4703, Beijing, China, October.

Sabrina Wilske. 2006. Dispositions for sociable robots. Master's thesis, Department of Computational Linguistics, Saarland University.

## Biographical Sketch

Sabrina Wilske is a PhD candidate within the International Research Training Group at Saarland University, Germany since May 2007. Her thesis is supervised by Prof. Manfred Pinkal and Magdalena Wolska (both Computational Linguistics, Saarbrücken). In May 2006 she earned a masters degree in Computational Lingustics from Saarland University, her master's thesis *Dispositions for Sociable Robots* was supervised by Prof. Hans Uszkoreit and Dr. Geert-Jan Kruijff (both German Research Centre for Artifical Intelligence). During her undergraduate studies she worked as a research assistant in the CoSy-project (Cognitive Systems for Cognitive Assistants) dealing with the natural language interface on robots.

[1] www.ltg.ed.ac.uk/dipper

[2] www.ling.gu.se/projekt/trindi/trindikit

[3] www.opendialog.org

**Craig Wootton**

University of Ulster
Shore Road
Belfast
Northern Ireland

[wootton-s1@ulster.ac.uk

## 1 Research Interests

My research interest lies in the area of **dynamic** dialogue systems, in particular investigating the feasibility and usability of **utilizing online content** for dialogue.

### 1.1 Past, Present and Future Work

Initial dialogue systems deployed in industry can be thought of as a rudimentary, but an effective, form of interaction, allowing customers to complete specific task based dialogues. Currently, dialogue systems are produced for specific tasks and domains.

Dialogue research meanwhile endeavors to advance these initial dialogue systems with the development of more natural and flexible systems, which can offer to the user more advanced features of dialogue, such as adaption to meet particular needs, evolving 'on-the-fly' dialogues to create more dynamic dialogues, or the engagement in dialogue in the true meaning of the word to help the user accomplish a particular task.

For these advanced features of dialogue systems to function, it is often the case that the dialogue knowledge and domain knowledge be separated from one another, and the domain knowledge represented in a well structured manner that is accessible to the dialogue manager. Structures like databases or ontologies, and associated query languages, are often used to store the domain knowledge, and the dialogue manager can simply query this knowledge representation when and how it needs to.

Contrast to this is the Internet, and the largely unstructured nature of the documents that make up the Internet, together with the vast range of topics available online. These have been a major issues and challenges for dialogue researchers wishing to complement the graphical browser with a dialogue interface to the online content. Current solutions have been to limit the dialogue manager's access and knowledge to a particular web site, reiterating the need for specific content types, sources and structures currently used in dialogue systems (Pargellis, Kuo & Lee 2004, Polifronti, Chung & Seneff 2003).

Another aspect of the problem is usability, which has been around for many years during the development of the computer and more so since the evolution of the graphical interface. Unlike their graphical counterparts, dialogue systems present a number of additional challenges for usability engineers. Inputs have to be constrained as per the language model, outputs needs to be relevant and meaningful to the user without being cognitively unmanageable, all the functionality of the system need to be obvious to the user, and error and confirmation strategies must provide an easy way for the user to recover in the events of miss and non understandings. Already challenging in the traditional 'task-based' dialogue systems, the issues are somewhat more complex when there is no set path through the dialogue or task to complete, or when the user has the initiative during the interaction. Due to the younger nature of usability research with regard to spoken dialogue systems, there is a current lack of usability studies in this 'informative' type of dialogue.

To combat and explore both the technical and usability issues presented, VoiceBrowse has been designed and implemented to further the work in this area of browsing the Internet through voice, not reliant upon specific content types or sources.

It has been proposed that **RSS and API feeds** can be used to gather content from the various online sources, building up a document source of available contents. This provides an XML based standardized *bridge* of accessing content and knowledge from any number of nonstandard sources. APIs from various providers, such as Amazon or Yahoo Travel, will provide a mechanism for driving *task based* dialogues, whereas the RSS feeds will facilitate *informative* dialogues based on content such as news items or weather. Once identified from the RSS feed, relevant content can be extracted from the associated webpage.

The VoiceBrowse architecture includes many different managers interacting, including *content spotter* being one novel component. Comparable to the *domain spotter* of the Queen's communicator (O' Neill et al. 2005) and also *evaluators* of the JASPIS architecture (Turunen et al. 2005), it provides a method of choosing the most relevant source of content based upon the user's query. Creating a document space from all the documents gathered from the RSS feeds, a cosine similarity function selects the best matched document matching the user's input query. The system has been developed to accommodate everyday tasks one would normally do online, such as request the news, book a flight, check the weather etc., and not as a question-answering tool.

To investigate the usability of such a system, two different implementations of VoiceBrowse have been created, a closed 'system led' initiative version, and an open 'user led' initiative version. It is the thesis that different users, of different ages, sex and technological experience will have different needs with respect to spoken dialogue usability, and will therefore interact with the two systems differently. The data collected during the evaluation on 32 participants will help further the work in this area of designing generic dialogue managers to interact with various content types.

## 2 Future of Spoken Dialog Research

Recently the context of the work of VoiceBrowse has been reinforced by the deployment of related 'Voice Search' related products, such as those made available from Microsoft and Yahoo. Like current dialogue systems deployed in industry, there is much research and development still required if this statement is to be realized, and if dialogue systems are indeed to offer natural, but robust, interfaces to interacting with machines generically, solving unrestricted problems and tasks.

Furthermore, if the goal of dialogue research is to mimic human interaction in its naturalness, then emotion detection will be required to be fused with dialogue research, considering human communication is often expressed through *how* something is said, rather than *what* is said. Consider, for example, a young baby about to touch something hot, to which the mother alerts the child by shouting warningly "Don't touch that!" The child of course is too young to understand the semantics or meaning of the words, but knows from both tone and characteristic of the mother's voice not to continue the movement of reaching out to touch the hot object. In the same way dialogue research should take into considerations, not only language understanding and linguistic analysis, but also efforts investigating the detection of emotion in speech, fusing the two inputs so a more accurate representation of the user's utterance.

Research efforts must also be concentrated on developing advanced standards for research, with which dialogue systems with sophisticated features can be developed by different parties using a common language. This would promote the collaboration of parties, leading to more effective research with a common direction and vision of dialogue systems.

## 3 Suggestions for discussion

- Do statistics speak for themselves? Will the statistical approach to language understanding completely eradicate the finite state grammar?
- Level playing field – can a standard set of metrics be devised to compare dialogue systems?
- GUIs killed spoken dialogue? Have GUIs evolved into such a usable product that dialogue systems are no longer seen as the usable natural interface they were once promised to be?

## References

O' Neill, I., Hanna, P., Liu, X., Greer, D. & McTear, M. 2005, "Implementing advanced spoken dialogue management in Java", *Science of Computer Programming,* vol. 54, no. 1, pp. 99-124.

Pargellis, A.N., Kuo, H.-J. & Lee, C.-. 2004, "An automatic dialogue generation platform for personalized dialogue applications", *Speech Communication,* vol. 42, no. 3-4, pp. 329-351.

Polifronti, J., Chung, G. & Seneff, S. 2003, "Towards the automatic generation of mixed-initiative dialogue systems from web content", *Eurospeech 2003*.

Turunen, M., Salonen, E., Hakulinen, J., Kanner, J. & Kainulainen, A. 2005, "Mobile Architecture for distributed multimodal dialogues", *Interspeech 2005* .

## Biographical Sketch

Craig Wootton is a PhD student at the University of Ulster under the supervision of Professor Michael McTear. He previously obtained his BSc in Computer Science at honors level. Currently in his final year of PhD study, research interests include the dynamic creation of dialogue, usability considerations of informative dialogues and investigating the generic nature of dialogue for reusing online content. Besides study, Craig's interest includes sport, music and film, and he also volunteers at his local Church where he leads various activities ranging from youth to musical groups.