

Toward Human-Like Task-based Dialogue Processing for HRI

Matthias Scheutz*, Rehj Cantrell** and Paul Schermerhorn**

Human-Robot Interaction Laboratory
*Tufts University and **Indiana University
Medford, MA, and Bloomington, IN, USA

Abstract

Many human social exchanges and coordinated activities critically involve dialogue interactions. Hence, we need to develop natural human-like dialogue processing mechanisms for future robots if they are to interact with humans in natural ways. In this paper, we discuss the challenges of designing such flexible dialogue-based robotic systems. We report results from data we collected in human interaction experiments in the context of a search task and show how we can use these results to build more flexible robotic architectures that are starting to address the challenges of task-based human-like natural language dialogues on robots.

Interactions in natural language dialogues are an essential part of human social exchanges, ranging from social conventions such as greetings, to simple question-answer pairs, to task-based dialogues for coordinating activities, topic-based discussions, and all kinds of more open-ended conversations. As a result, the ability of future social and service robots to interact with humans in natural ways (Scheutz et al. 2007) will critically depend on developing capabilities of human-like dialogue-based natural language processing (NLP) in robotic architectures. However, different from other NLP contexts such as story understanding or machine translation, natural language processing on robots has at least the following six properties: *real-time*, *parallel*, *spoken*, *embodied*, *situated*, and *dialogue-based*.

Real-time means that all processing must occur within the time frame of human processing, both at the level of comprehension as well as production. It also means that constraints will have to be incorporated incrementally as they occur, analogous to human language processing.

Parallel means that all stages of language processing must operate concurrently to mutually constrain possible meaning interpretations and to allow for the generation of responses (such as acknowledgements) while an ongoing utterance is being processed.

Spoken means that language processing necessarily operates on imperfect acoustic signals with varying quality that depends on the speaker and the background noise. In addition to handling prosodic variations, this includes typical features of spontaneous speech such as various types of dis-

fluencies, slips of the tongue, or other types of errors that are usually not found in written texts.

Embodied means that robots have to be able to process multi-modal linguistic cues such as deictic terms accompanied by bodily movements, or other gestures that constrain possible interpretations of linguistic expressions. It also means that the robot will have to be able to produce similar gestures that are expected by human interlocutors to accompany certain linguistic constructs.

Situated means that, because speaker and listener are located in an environment, they will have a unique perspective from which they perceive and experience events, which, in turn, has an impact on how sentences are constructed and interpreted. This includes the incremental integration of perceivable context in the interpretation of referential phrases as well as being sensitive to non-linguistic coordination processes such as the establishment of joint attention.

Dialogue-based means that information flow is not unidirectional, but includes bidirectional exchanges between interlocutors based on different dialogue schemes which constrain the possible dialogue moves participants can make at any given point.

While the above six aspects present significant challenges for the development of robotic architectures with dialogue capabilities, there are also several advantages to natural language processing on robots that other NLP contexts do not have. For example, spoken natural language exchanges typically consist of shorter sentences with usually simpler grammatical constructions compared to written language (thus making parsing easier and more efficient). Moreover, the employed vocabulary is much smaller and the distribution of sentence types is different (including more commands and acknowledgements, and few declarative sentences compared to written language). Also, different from written texts, perceptual context can be used to disambiguate expressions, and most importantly, ambiguities or misunderstandings in general can often be resolved through subsequent clarifying dialogue. The option to request clarification also allows interlocutors to handle new, unknown expressions naturally.

Since there are many different forms of dialogues that have their own rules and conventions based on social norms and etiquette (such as small talk, interviews, counseling talks, etc.) and might, moreover, require tracking of various non-linguistic aspects (such as contextual information,

interlocutor eye gaze and affective as well as other mental states), we focus on task-based dialogues in the paper. We start by briefly reviewing the shortcomings of current dialogue-based NLU systems on robots, and then discuss our attempts to address some of them. We next report results from human experiments that we conducted to collect a dataset as the basis for the architecture development and then give examples of the kinds of dialogue exchanges our current system can handle.

Previous Work

Although rapid progress in this field of research has been made recently, no current systems are yet capable of these types of dialog interactions. Moreover, most robotic systems lack one or more of the other essential requirements such as incrementality of processing and real-time operation, the inclusion of contextual constraints on semantic processing, the usage of a training grammar rather than an exclusively rule-based grammar (created laboriously by hand), and mechanisms for handling disfluencies and other distortions of spontaneous speech.

For one, robotic systems that use natural language for robot instruction either do not have natural language fully integrated into the robotic architecture (e.g., (Michalowski et al. 2007)) or are limited to simple instructions (e.g., (Firby 1989; Atrash et al. 2009)). And many systems use rule-based grammars due to the difficulty of producing a well-trained grammar on what is invariably the small amount of data applicable to the domain. For one example, (Bos and Oka 2007) uses Nuance to produce first-order λ -expressions according to a rule-based grammar. Disfluencies are not allowed, and the system is nonincremental, in that it must process the full utterance before proceeding further.

While there are several examples of incremental systems (e.g., (Winograd 1972; Erman et al. 1980; Lowerre and Reddy 1980) and more recently (DeVault and Stone 2003; Allen et al. 1996; Varges and Purver 2006)), they usually only process linguistic information. An exception is the system by Schuler et al. (Schuler, Wu, and Schwartz 2009) which integrates phonological, syntactic, and referential semantic information into a single language model, thus enabling semantic information to boost probabilities for word hypotheses of existing objects in the space. However, the system has not been used on a robot and would require a large amount of training data (often not available) in order to be able to extract the relevant statistical information. Other complete NLU systems that do not require large data sets for training (e.g., TRIPS (Allen, Swift, and de Beaumont 2008)) are typically not incremental and have usually not been designed with the challenges of NLP on robots in mind.

One approach to improving robustness is to allow for partial parses. For example, (Rickert et al. 2007) uses OpenCCG, a rule-based, non-incremental parser, to produce an unconstrained λ -logical semantic representation of an input. When it is not able to form a complete semantic parse, it produces partial parses for further processing, rather than failing outright. In (Atrash et al. 2009), disfluencies are not specifically handled, but if a parse fails, the speech recognizer output is passed to the goal manager as a

bag of words. (Kacalak and Majewski 2009) extracts known words into commands rather than attempting to create a full parse from input that may also contain unknown words, and the word-recognition module allows malformed words to be corrected.

Human Performance in a Search Task

It is difficult, if not impossible, to anticipate *a priori* the wide variety of spoken task-based natural language interactions that will be initiated by humans during interactions with robots, even when the task is well-defined. Hence, we conducted several human-human studies in a task where two humans must coordinate their activities via remote audio communication only (i.e., no visual links); their transcribed interactions form the “Cooperative Remote Search Task” (CReST) corpus (Eberhard et al. 2010; Scheutz and Eberhard 2008a), which we use to identify natural interaction patterns as well as potential pitfalls.

Dialogue-based natural language interactions are quite different from natural language instructions as they would be given in written form. For example, consider a written instruction for following a particular route such as “Continue to walk straight, going through one door until you come to an intersection just past a whiteboard.” (Kollar et al. 2010). An interactive version of the same instruction from the corpus is substantially more complex:

Instructor: OK, continue to walk straight.

Robot (continuing straight): OK.

Instructor: You should be seeing a door in front of you.

Robot (looking out for a door): Yes.

Instructor: Good, go through that door.

Robot (moving through the door): OK, I’m through the door.

Instructor: Alright. Keep going. There should be a whiteboard.

Robot (looking for whiteboard): OK, I’m not seeing it yet. Ah, there it is.

Instructor: Great, then you should see an intersection, go there.

Robot (looking out for an intersection while moving): Got it, OK.

Note that natural language instructions are given piecemeal with the expectation of rapid feedback, and meanings as well as goals are “negotiated” through a sequence of dialog moves and actions, rather than being fixed in a sequence of instructions. As a result, perception, NLU, and behavior have to be tightly intertwined. It is also expected that any possible ambiguities (e.g., the presence of multiple doors) can be resolved quickly online, hence there is no need to provide abundant, fully self-contained information; rather humans provide minimal information, happy to refine it or add to it during the ensuing dialog-interaction. Among the most frequently-occurring issues identified in the corpus¹ are *ungrammatical sentences*, including incomplete referential phrases, missing verbs, corrections, and others; *wrong*

¹The CReST corpus, including transcriptions of all interactions coded for disfluencies and dialogue moves, together with POS tags and dependency and CCG parse annotations, will be made available for research purposes free of charge.

word substitutions for intended target words such as “block” and “book” for “box”; *underspecified directions, referents, and directives* which assume shared task-knowledge, knowledge of subgoals, perspectives, etc.; *frequent “ums”, “uhs”, and other disfluencies and pauses* indicating cognitive load; and *frequent coordinating confirmations and acknowledgments* as dialogue moves including prosodically different “okays”, “yeahs”, etc.

Despite attempts to address individual challenges of robot NLP, no existing system successfully manages the *real-time, parallel, spoken, embodied, situated, and dialogue-based* NLP described above. Almost all current robotic architectures suffer from insufficient NLP speed, inability to systematically handle the disfluencies that come up naturally in spontaneous speech, and from limitations on the dialogue side, including feedback via gestures and integration of dialogue schemes to ensure appropriate feedback is given.

Tackling Task-Based Dialogue HRI

The dialogues in the corpus are typical examples for the kinds of coordinating natural language interactions humans exhibit in limited domains like cooperative search tasks. The ten most frequent word types in the entire corpus (excluding acknowledgments) are the, box, I, a, 's, there, you, on, and, and that. The 151 most frequent words cover 60% of the tokens, 280 words reach 80% coverage, and 712 reach complete coverage. Yet, despite their seemingly very limited nature (based on vocabulary), the dialogues present a major challenge for HRI. It is clear that meanings are not constructed from sentences alone but from interactions that serve particular purposes and accomplish particular goals. Perception, action, and language processing in humans are obviously all intertwined, involving complex patterns of actions, utterances, and responses, where meaningful linguistic fragments result from their context together with prosodic, temporal, task and goal information, and *not* sentence boundaries. Consequently, we need to develop new models of interactive natural language processing and understanding for HRI that process language in very much the same interactive, situated, goal-oriented way as humans. In particular, we need to integrate the timing of utterances, back-channel feedback, perceivable context (such as objects, gestures, eye gaze of the participants, posture, etc.), as well as background and discourse knowledge, task and goal structures, etc., if we want to achieve human-level performance on robots. This poses both *functional challenges* and *architectural challenges*.

Functional challenges include (1) mechanisms for providing appropriate feedback that humans expect even while an utterance is still going on, using different kinds of acknowledgment based on dialogue moves; (2) new algorithms for anaphora and reference resolution using perceptual information as well as task and goal context; and (3) mechanisms for handling various kinds of disfluencies and incomplete and ungrammatical utterances, including robust speech recognition, parsing, and semantic analysis.

Architectural challenges include (1) real-time processing of all natural language interactions within a human-acceptable response time (e.g., typically acknowledgments

have to occur within a few hundred milliseconds after a request); (2) integration of various natural language processing components (including speech recognition, parsing, semantic and pragmatic analyses, and dialogue moves) that allows for parallel execution and incremental multi-modal constraint integration; and (3) automatic tracking of dialogue states and goal progress to be able to provide meaningful feedback and generate appropriate goal-oriented dialogue moves.

While each of the above functional challenges is a research program in its own right and we are nowhere close to addressing any of them in a satisfactory manner, it is still possible to make progress in parallel on the architectural challenges, i.e., on defining appropriate functional components in the architecture with appropriate data structures and information flow among them to facilitate the integration of the algorithms that will meet the functional challenges.

Over the last decade, we have started to address natural language dialogue interactions on robots in our *distributed, integrated, affective, reflective, cognitive* (DIARC) architecture (Scheutz et al. 2007). DIARC is implemented in the *agent development environment* (ADE), a framework for implementing architectures that provides infrastructure support for deploying functional components of the architecture across multiple hosts, discovery of available services (provided by other components), and mechanisms for improved reliability and, when necessary, recovery from component failures. The DIARC architecture features tight multi-level integration of goal management and action selection with low-level sensor and effector components providing access to a wide variety of hardware (mobile robots, humanoid torsos, cameras, 3D imagers, microphone arrays, etc.). The priority-based goal manager supports concurrent pursuit of multiple goals in independent action script interpreters, as long as there are no conflicts; when two goals require access to the same resource, the conflict is resolved in favor of the one with the highest priority.

Of greatest relevance to the present context, however, are DIARC's natural language processing capabilities. For example, we developed and implemented algorithms for human-like incremental reference resolution on a robot (Scheutz, Eberhard, and Andronache 2004), which we subsequently extended to allow for dialog-like HRI with simple forms of backchannel feedback such as nodding or saying “okay” (Brick and Scheutz 2007). We also integrated NLP components tightly with action execution (Brick, Schermerhorn, and Scheutz 2007), a pre-requisite for the robot's ability to start actions quickly (e.g., nodding). More recently, we demonstrated algorithms for automatically converting natural language instructions into formal *goal interpretation* (expressed in a fragment of the “computational tree logic” CTL*) and *action interpretation* for achieving the goal (expressed in propositionalized first-order form in “propositional dynamic logic” PDL). The conversion into logical forms is effected by combining lexical items with syntactic annotations from a combinatorial categorial grammar (CCG) and semantic annotations from the two logics extended by λ -expressions. Repeated λ -conversions then lead to λ -free temporal and dynamic formulas that represent

the goals and actions specified in the natural language instruction, respectively. We also have developed algorithms for handling disfluencies, in particular, lexical disfluencies, abandoned utterances, repetitions, as well as some repairs and corrections, in the context of spoken instruction understanding (Cantrell et al. 2010). Much of the above work was directly based on our CReST corpus.

We will now present three example dialogues to illustrate different aspects of typical human-like dialogues currently possible in our DIARC architecture (Scheutz et al. 2007). All dialogues have been performed on different types of robots.

Handling Disfluencies

There are several basic types of disfluencies in CReST: repetitions, insertions, abandoned utterances, and repairs.

Repetitions of exact words or word sequences may be one word:

Director: so two doorways and then
you'll you'll be staring straight
at a platform

or several words in length:

Director: is that a new is that a new green
box that you didn't tell me about

Insertions may be words (lexical) or nonwords (nonlexical). Nonlexical insertions include, e.g. "uh", "um". Lexical insertions can be similar to repetitions but are not exact:

Director: how many box how many blue
boxes do we have

Repairs, a subclass of insertions, denote instances in which one word is replaced by another. This example shows a simple one-word correction – replacing "at" with "by":

Searcher: one green box at the corner at
by the end of the hallway

In the next example, an *aborted word* (i.e. a nonword) is inserted:

Searcher: and gre- green box in the right
hand cubicle

Abandoned utterances may or may not be followed immediately by a distinct sentence. In the first example, the speaker abandons an utterance and does not start a new one:

Director: so pink boxes should uh only
be in -

while in the next example an utterance was abandoned and a new one was started:

Searcher: so to my - I'm just like next to
the door frame like right in front
of it and there's a desk to my
left

Although lexical insertions and abandoned utterances can seem similar, insertions tend to be similar to the following phrase, whereas abandoned utterances tend to be followed by a completely separate thought.

In somewhat rarer cases, an entirely wrong word is used:

Director: and then there should have been a
blue box by the cardboard box at
the end of the hallway but you
said there wasn't so that's the
misplacement that they spoke of
earlier

Searcher: I didn't see one at the end of
that hardware

Director: yeah there definitely wasn't
blue one at the end of that
hallway with the cardboard box

Surrounding context clearly indicates that "hardware" referred to a "hallway", yet the misspeak is never acknowledged (and possibly never noticed) by either participant.

Rule-based algorithms for each of these disfluency types have shown some success. However, these methods are typically targeted at offline NLP, and rely on transcription features that are not usually provided by real-time speech recognizers, such as identifying aborted words by looking for those that end with hyphens. Our method, given a possibly disfluent utterance, is to produce only a partial parse with whatever can be made sense of, similar to other robotic systems noted above. Having been trained on correct (i.e. non-disfluent utterances), the parser, when faced with disfluent utterances, may do any of several things in order to discard such disfluencies:

Attach extra nodes to the root. Each root node is viewed as a separate phrase; attaching disfluent words to the root creates semantically incomplete phrases that are subsequently discarded. In "to go to go into a room", the disfluent initial nodes "to go" were both left unconnected to the rest of the graph, leaving only the relevant parts of the utterance. This is shown in Figure 1(a).

Attach extra nodes to a nearby node. Given the connection rules, this will often result in correct semantic output. For example, "you are *um* at a closed *um* door" should, regardless of the nonlexical fillers, result in $at(listener, x)$ where x is known to be a door that is closed. Figure 1(b) shows how this works for the first occurrence of "um": the parser connected "at" as a *pred* child of "um", which in turn was listed as a *pred* child of "are". The definition of "are" indicated that it should take a predicate and a subject, and attach the subject to the predicate. The system created an empty definition for "um" which allowed it to attach the subject it was handed by "are" to its own predicate child "at", resulting in the correct definition.

Leave extra nodes unattached. In Figure 1(b) the second "um", between "a" and "closed", was simply not attached to any node and was thus discarded.

Note that in none of these cases is it necessary to explicitly identify the disfluency. Instead, these actions are chosen by the parser as it attempts to incrementally parse the input. With well-formed input, the same rules lead to correct parse trees, whereas disfluencies end up "falling out" of the parse as a side effect of the parsing process, as described above.

Handling multi-modal natural spoken dialogues

The next example is an interaction that takes place between an autonomous mobile robot with a humanoid torso and a human interlocutor. They engage in a simple dialogue that explores the robot's perceptual capabilities (such as its ability for perspective taking). In this case, it is important that acknowledgments include both verbal and non-verbal components, because the robot's humanoid features trigger expectations that it will behave in a human-like way, using multiple channels to convey information (e.g., producing head motions in conjunction with utterances to indicate agreement or disagreement).

Human: Hello robot.

Robot: Hello.

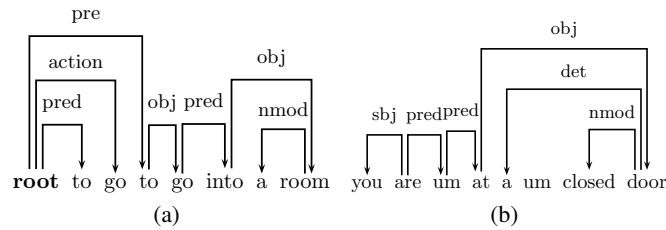


Figure 1: Graphical examples of disfluent dependency graphs. Arcs represent head-child relations, with the arrow pointing from the head to the child. Children of the *root* node are considered to be heads of distinct phrases or utterances.

Human: Come over here.

The robot uses sound source localization to determine the location of the human speaker to resolve the indexical “here” as part of the action *move-to(here)* (i.e., the meaning of “here” in this context is the location of the speaker).

Human: Do you see a green, no, red box?

Robot: (shakes head) No, I don’t see anything red.

The robot handles the repair “green, no, red” correctly and searches for a red box. Since it does not see a red box, and, in fact, sees nothing red, it generates a head motion together with an utterance stating as much.

Human: Look to the left.

Robot: (looks to its left)

Human: Now do you see it?

Robot: (nods head) Yes, I see a red one.

The robot takes the instruction to look to the “left” to mean “left from its perspective” and turns its head in that direction.² It ignores the word “now” at the beginning of the following sentence and resolves the pronoun reference of “it” to refer to the “red box” from the previous sentence. Upon noting the red box now in its visual field, it nods and generates its own anaphoric reference using “one” instead of “box”.

Human: Is the red box to the left of the blue box?

Robot: (nods head) Yes, it is.

Human: Well, ah, is it to the right of the blue box?

Robot: (shakes head) No, it is not.

Again, the robot handles anaphoric reference correctly and also handles fillers (“well”) and disfluencies (“ah”).

Human: Point to the blue box.

Robot: (points to the blue box) It’s over there.

Human: (moves out of robot’s visual field) Look at me.

Robot: (looks at the human)

Human: Thank you.

Robot: (smiles) You’re welcome.

The robot accompanies its pointing action with verbal feedback using the indexical “there”, which in this case indicates the location of the pointing action. Moreover, it resolves the indexical reference of “me” in “look at me” and

uses sound localization again to resolve the destination of the “look” command (i.e., the location of the speaker, who has moved out of the robot’s field of vision – had the speaker remained within the robot’s visual field, the robot could have used vision to orient its head to face the speaker). See <http://www.youtube.com/watch?v=NinDDNc7sCM>.

Handling dialogues with task-based knowledge

The next example also builds on the search task (Scheutz and Eberhard 2008b) and demonstrates the kinds of natural task-based dialogues found in the human experiments. In particular, the example shows how the robot can determine what objects to look out for and what facts about the environment to report based on its knowledge of the task. It also demonstrates how the robot can quickly react to a given instruction by changing its behavior based on a tight integration between natural language understanding and action execution (Brick, Schermerhorn, and Scheutz 2007) while providing the necessary feedback to the human commander. At this point in the interaction, the robot has been going down one corridor and has stopped outside of a doorway:

Human: Is there a hallway?

Robot: I see a hallway.

Human: Okay, go down there.

Robot: Okay.

The robot drives down the newly-encountered hallway. As it is driving down it notices a doorway, which it reports to the team leader, also acknowledging its position:

Robot: Okay, I’m now in the hallway.

There is a doorway on the left.

Human: Good, go through that doorway.

Robot: Okay.

The robot enters the room and notices several yellow blocks, some of which are in boxes. Since these are task-relevant, it reports them to the team leader:

Robot: I’m now inside the room. There are yellow blocks in boxes.

Human: Get a yellow block from a blue box.

Robot: Okay.

The robot verifies that there is a yellow block in a blue box, approaches it, and gets it from the box. See the video at <http://www.youtube.com/watch?v=Lr3pNDJ3XIA>.

Discussion

While the above examples demonstrate some encouraging progress, there is still a long road ahead with many obstacles that need to be overcome before robots will be able to

²Of course, it is possible that the interlocutor had a different perspective in mind; the same is possible, of course, in human-human interactions, and may require subsequent synchronization of the two agents’ assumptions.

engage in human-like natural language dialogues. Some of the obstacles such as speech recognition have been known for a long time. After years of significant efforts and great progress, there is unfortunately still no speech recognizer that can provide sufficiently high recognition rates in typical human environments that are not carefully controlled for background noise. Similarly, most current parsers need large amounts of training data to be able to perform well on novel problems. However, such training data is usually not available for HRI domains and collecting it is often not feasible (and even when it is feasible, it is unclear how much of the data will transfer to different tasks in different environments). There are also open questions about dialogue systems that are currently not fully understood, for example, the different types of dialogue moves in task-based dialogues and the resultant interaction schemes that humans will naturally follow. Moreover, it is currently unclear to what extent robots will have to build and maintain mental models of their human interlocutors to be able to reach human-like dialogue competence. Beyond the above linguistic issues, there are other important open questions related to a robot's physical appearance and capabilities. It is, for example, unclear whether people will even seriously engage in human-like dialogues with robots that have very different physical forms (e.g., wheels, no heads, etc.) and very different physical capabilities. And it is an open question how super-human capabilities (e.g., being able to follow tens of conversations at the same time) would be received should those become possible at some point. Clearly, additional HRI experiments will be needed to stake out the territory of human-robot dialogue, human receptivity to it, and the best ways for robots to meet human expectations.

Conclusion

In this paper, we demonstrated the challenges faced by a robotic natural language understanding system that is intended for natural human-like dialogue interactions with the robot in HRI contexts. And we briefly introduced our DIARC architecture which is starting to address some of these challenges. Specifically, we reported results of several simple natural dialogues that DIARC can handle together with a brief high-level description of the processes involved in handling them. Clearly, this is only a start and there is a great deal of work ahead of us before we can claim to have reached *natural* human-like spoken language interactions. However, by starting with limited domains and tasks (such as instruction tasks), it is possible to make progress right now toward a not-too-distant future point where the resultant architecture will be ready for transition into real-world application domains.

Author Bios

Matthias Scheutz received degrees in philosophy (M.A. 1989, Ph.D. 1995) and formal logic (M.S. 1993) from the University of Vienna and in computer engineering (M.S. 1993) from the Vienna University of Technology (1993) in Austria. He also received the joint Ph.D. in cognitive science and computer science from Indiana University in 1999.

Matthias is currently an associate professor of computer and cognitive science in the Department of Computer Science at Tufts University. He has over 100 peer-reviewed publications in artificial intelligence, artificial life, agent-based computing, natural language processing, cognitive modeling, robotics, human-robot interaction and foundations of cognitive science. His current research and teaching interests include multi-scale agent-based models of social behavior and complex cognitive and affective robots with natural language capabilities for natural human-robot interaction.

Rehj Cantrell received degrees in Physics (B.S., 2003, Purdue University) and Computational Linguistics (B.A. 2008, M.A. 2011, Indiana University). She is currently a Ph.D. student at Indiana University. Her research interests are natural language understanding for human-robot interaction, dependency parsing, and dialogue systems.

Paul Schermerhorn received degrees in Philosophy (M.A. 1999, Northern Illinois University) and Computer Science (M.S. 2002, Ph.D. 2006, University of Notre Dame). He is currently an Assistant Research Scientist in the School of Informatics at Indiana University, Bloomington. He has over 10 years of research experience, and over 40 peer-reviewed publications, in artificial life, robotics, and human-robot interaction.

Acknowledgments

This work was in part funded by ONR MURI grant #N00014-07-1-1049 to the first author.

References

- Allen, J. F.; Miller, B. W.; Ringger, E. K.; and Sikorski, T. 1996. A robust system for natural spoken dialogue. In Joshi, A., and Palmer, M., eds., *Proceedings of the Thirty-Fourth Annual Meeting of the Association for Computational Linguistics*, 62–70. San Francisco: Morgan Kaufmann Publishers.
- Allen, J.; Swift, M.; and de Beaumont, W. 2008. Deep semantic analysis of text. In *Proceedings of the Symposium on Semantics in Systems for Text Processing (STEP)*.
- Atrash, A.; Kaplow, R.; Villemure, J.; West, R.; Yamani, H.; and Pineau, J. 2009. Development and validation of a robust speech interface for improved human-robot interaction. *I. J. Social Robotics* 1(4):345–356.
- Bos, J., and Oka, T. 2007. A spoken language interface with a mobile robot. *Artificial Life and Robotics* 11:42–47.
- Brick, T., and Scheutz, M. 2007. Incremental natural language processing for HRI. In *HRI*, 263–270.
- Brick, T.; Schermerhorn, P.; and Scheutz, M. 2007. Speech and action: Integration of action and language for mobile robots. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*.
- Cantrell, R.; Scheutz, M.; Schermerhorn, P.; and Wu, X. 2010. Robust spoken instruction understanding for hri. In *HRI 2010*, (forthcoming).
- DeVault, D., and Stone, M. 2003. Domain inference in incremental interpretation. In *Proc. ICoS*.

- Eberhard, K.; Nicholson, H.; Kuebler, S.; Gundersen, S.; and Scheutz, M. 2010. The indiana cooperative remote search task (crest) corpus. In *Proceedings of LREC 2010: Language Resources and Evaluation Conference*.
- Erman, L. D.; Hayes-Roth, F.; Lesser, V. R.; and Reddy, D. R. 1980. The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys* 12(2):213–253.
- Firby, R. J. 1989. *Adaptive Execution in Complex Dynamic Worlds*. Ph.D. Dissertation, Yale University.
- Kacalak, W., and Majewski, M. 2009. Natural language human-robot interface using evolvable fuzzy neural networks for mobile technology. In Huang, D.-S.; Jo, K.-H.; Lee, H.-H.; Kang, H.-J.; and Bevilacqua, V., eds., *Emerging Intelligent Computing Technology and Applications*, volume 5754 of *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. 480–489. 10.1007/978-3-642-04070-2_54.
- Kollar, T.; Tellex, S.; Roy, D.; and Roy, N. 2010. Toward understanding natural language directions. In *HRI '10: Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, 259–266. New York, NY, USA: ACM.
- Lowerre, B., and Reddy, R. 1980. The Harpy speech understanding system. In Lea, W. A., ed., *Trends in Speech Recognition*. Prentice Hall. 340–360.
- Michalowski, M. P.; Sabanovic, S.; DiSalvo, C.; Font, D. B.; Hiatt, L.; Melchoir, N.; and Simmons, R. 2007. Socially distributed perception: GRACE plays social tag at AAAI 2005. *Autonomous Robots* 22(4):385–397.
- Rickert, M.; Foster, M. E.; Giuliani, M.; By, T.; Panin, G.; and Knoll, A. 2007. Integrating language, vision and action for human robot dialog systems. In *Proceedings of the 4th international conference on Universal access in human-computer interaction: ambient interaction*, UAHCI'07, 987–995. Berlin, Heidelberg: Springer-Verlag.
- Scheutz, M., and Eberhard, K. 2008a. Towards a framework for integrated natural language processing architectures for social robots. In *Proceedings of the Fifth International Workshop on Natural Language Processing and Cognitive Science (NLPCS-2008)*.
- Scheutz, M., and Eberhard, K. 2008b. Towards a framework for integrated natural language processing architectures for social robots. In *Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science*, 165–174.
- Scheutz, M.; Schermerhorn, P.; Kramer, J.; and Anderson, D. 2007. First steps toward natural human-like HRI. *Autonomous Robots* 22(4):411–423.
- Scheutz, M.; Eberhard, K.; and Andronache, V. 2004. A real-time robotic model of human reference resolution using visual constraints. *Connection Science Journal* 16(3):145–167.
- Schuler, W.; Wu, S.; and Schwartz, L. 2009. A framework for fast incremental interpretation during speech decoding. *Computational Linguistics* 35(3):313–343.
- Varges, S., and Purver, M. 2006. Robust language analysis and generation for spoken dialogue systems. In *Proceedings of the ECAI workshop on Development and Evaluation of Robust Spoken Dialogue Systems for Real Applications*.
- Winograd, T. 1972. *Understanding Natural Language*. Academic Press.