Artificial Intelligence for Human-Robot Interaction
Papers from the AAAI 2015 Fall Symposium

Agent Requirements for Effective and Efficient Task-Oriented Dialog

Shiwali Mohan[*], James Kirk[**], Aaron Mininger[**], John Laird[**]
[*]PARC, 3333 Coyote Hill Road, Palo Alto, CA 94305
[**]University of Michigan, 2260 Hayward Ave., Ann Arbor, MI 48109-2121
shiwali.mohan@parc.com, jrkirk@umich.edu, mininger@umich.edu, laird@umich.edu

Abstract

Dialog is a useful way for a robotic agent performing a task to communicate with a human collaborator, as it is a rich source of information for both the agent and the human. Such task-oriented dialog provides a medium for commanding, informing, teaching, and correcting a robot. Robotic agents engaging in dialog must be able to interpret a wide variety of sentences and supplement the dialog with information from its context, history, learned knowledge, and from non-linguistic interactions. We have identified a set of nine system-level requirements for such agents that help them support more effective, efficient, and general task-oriented dialog. This set is inspired by our research in Interactive Task Learning with a robotic agent named Rosie. This paper defines each requirement and gives examples of work we have done that illustrates them.

Introduction

Task-oriented dialog between a human and robot consists of a series of interactions where the purpose is to transmit knowledge between agents to facilitate coordination, cooperation, and potentially learning in order to complete some task. For effective communication with a person, agent design should be motivated by human-human collaboration and learning, from the different modalities of interactions supported to how percepts and knowledge are referenced and shared. In human dialog interactions, natural language is often the medium of transmission, but it is often augmented by gestures or facial expressions. It makes use of symbolic references to objects, states, relationships, events, and collections thereof. In human collaborative teams, it is employed to convey many types of information including perceptual features relevant to performing a joint task, hierarchical task structures, and joint goals (Grosz & Sidner, 1986; Oviatt & Cohen, 1991; Bangalore et al. 2008; Scheutz, et al. 2011). In learning scenarios, it is employed to transfer information about the world, correct behavior (Litman & Allen, 1987), or to request and provide clarifications (Litman & Allen, 1990).

Unsurprisingly, several recent efforts (Matuszek et al., 2012; Kollar et al., 2010; Tellex et al., 2011) have pursued task-oriented dialog capabilities for collaborative robots. Most of these efforts have studied a sub-problem of task-oriented dialog – grounding linguistic symbols and constructions to sensory information and control structures, where the purpose is to command the robot to perform a specific task. However, in order to develop truly interactive robots that can effectively collaborate across multiple tasks and situations, several other aspects of task-oriented dialog need to be addressed. The main body of this paper identifies a set of nine system-level requirements for developing a task-oriented dialog capability, and illustrates them using examples drawn from our experience of developing interactive task learning agents (Kirk & Laird, 2014; Mohan et al. 2012; Mohan & Laird, 2014), which we briefly introduce in the next section. One of our conclusions is that online, interactive learning is important for realizing general, effective, and efficient task-orientated dialog on a robotic system.

Interactive Task Learning Agent: Rosie

The main thrust of our research in HRI has been on Interactive Task Learning (ITL: Laird, 2014), where a robotic agent learns not only the procedures for performing a task, but also the definitions of that task. Our system, named Rosie, is an example of such an agent. Rosie relies upon task-oriented dialog with a human instructor to acquire new knowledge. It learns many aspects of tasks, including the relevant objects and actions, task goals, execution policy, and preconditions for attempting the task. Rosie can also learn knowledge relevant to games (e.g., Tic-Tac-Toe), such as the rules of the game, constraints on actions (*you can't move a piece that has been played*), and winning and losing conditions (three-in-a-row). Although there are other means for teaching an agent a new task, task-oriented dialog is ubiquitous in human instruction and can be efficient, effective, and natural for humans to use (Mo-

han 2015). Rosie is implemented in Soar (Laird, 2012) and is embodied in a table-top robot. It also has recently been ported to a mobile robot where we are teaching it delivery, patrolling, and similar tasks.

Agent Requirements for Task-Oriented Dialog

Based on our experience with ITL and more specifically our implementation of Rosie, we have identified a set of requirements for supporting task-oriented dialog, driven by the following criteria:

- Generality: the agent supports many forms of dialog and modalities of interaction for different types of knowledge across multiple domains and tasks.
- Efficiency: the number of interactions and the length of the communications (number of words or gestures) required to transmit a concept is minimized.
- Effectiveness: the agent uses the transmitted information to direct its behavior and learn about its environment as well as produce useful information in its communications.

Encode a common reference scheme

Task-oriented dialog contains references to relevant aspects of the environment and the task. For effective communication, collaborators must have a common reference scheme that specifies how these references will be made. The agent must know how such references relate to what it can perceive on its sensors and what it knows about its environment. This knowledge may be pre-programmed or learned through experience. This is the grounding problem which has been widely studied in the prior work.

For Rosie, the instructor can identify objects using different references, including linguistic descriptions (*the red object*), gestures (indicating an object and using the word *this*), and spatial constraints (*the block on the blue rectangle*). In all cases, the reference is resolved to an internal symbolic representation of the object. Symbols referring to visual properties are also grounded in subsymbolic information. For example, *red* corresponds to a region of a feature space in a perceptual classifier. Spatial relations are grounded in comparisons of metric information about the objects (spatial comparisons of bounding boxes). Through mappings like these, Rosie grounds words and phrases to objects in the environment.

These mappings can be preprogrammed, but Rosie also learns concepts in order to handle novel environments and tasks. Rosie learns new nouns and adjectives, such as label names (*kitchen, pantry*), colors (*red, blue*), shapes (*rectangle, triangle*), and sizes (*large, small*), and prepositions (*on, left of).* Regardless of how each concept is taught, or through which modality of interaction, Rosie constructs a symbolic representation (*red1*) of the concept that connects to related symbolic knowledge (*type: color*), subsymbolic knowledge (RGB categorization space), and the linguistic term used in the dialog ("red"). Rosie learns similar groundings for the other concepts in order to maintain a uniform representation of the various types of knowledge that connect observations of the environment, references in the dialogue interaction, and previously learned knowledge.

Encode typical information-exchange protocols

In task-oriented dialog, prototypical types of utterances are employed for exchanging different types of information. Imperative sentences, such as *put that book in the shelf*, are used to covey an intended goal to be achieved by the listener. Assertions such as *there is a blue couch in the living room* are used to convey a belief about the environment. A question (*where is the milk?*) can be employed to supplement perceptual information by relying on the collaborative partner's knowledge of the environment. In order to correctly interpret or generate these utterances, the agent must understand how the structure of an utterance relates to the information that is being provided or requested.

Rosie has a simple referential grammar (implemented as rules) to generate reasonable responses to the instructor's utterances. This grammar assumes nouns (*block*) and adjectives (*red*) refer to visual properties of objects sensed through the camera; referring expressions (*the red large block*) refer to specific objects in the environment; prepositions (*behind*) refer to spatial relationships between objects; and verbs (*place*) refer to abstract task descriptions. Imperative sentences composed of these elements (*place the red large block behind the small yellow triangle*) refer to a task representation instantiated with relevant objects, relationships, and a goal. Comprehension of these imperatives results in Rosie executing the instantiated task in the environment. Rosie also understands and generates questions (*what kind of attribute is purple?*) and statements (*the red block is behind the yellow triangle*).

The implemented grammar is small and is designed with a perspective that Rosie functions in the ITL domain where its primary goal is to learn new tasks. While the grammar has been useful in teaching Rosie a variety of tasks, it is hand-engineered. An ideal learner should expand its knowledge of information-exchange protocols as it gathers interactive experience. Currently, Rosie can learn to ground new words to components of its environment and new verbs to actions, but it cannot learn new types of utterances and how they connect to its behavior and goals.

Incorporate non-linguistic context for interpretation

Situated dialog between humans is efficient – information that is apparent from the current environment or is a component of shared beliefs does not have to be explicitly identified or transmitted by the speaker. This may sometimes lead to ambiguity in the utterances employed by the speaker. The agent must use its knowledge and experience

of the environment to effectively interpret the dialog within the current context.

In Rosie, non-linguistic context is used for object reference disambiguation. A simple form is when human instructions use the referent *this*, accompanied by a gesture involving clicking on a simulated view of the object. For references such as *it*, Rosie uses the sentence structure as well as the event and dialogue history as context to determine the referred object. In other cases, Rosie uses the context of the current spatial scene to decide between ambiguous preposition attachments, incomplete referring expressions, and polysemous verbs. For example, "move the red block on the stove" in interpreted differently if the red block is already on the stove than if it is somewhere else.

Additional context about the current purpose of the interaction can aid interpretation, particularly for sentence fragments. For example, when Rosie encounters an unknown concept, *red,* it asks *what type of attribute is red?* The response *a color* can be understood in context of the question as equivalent to the fuller response *red is a color*.

Incorporate information from multiple interaction modalities

Natural language is not always the most efficient way to communicate knowledge to the agent. Human dialog is often augmented with gestures, facial expressions, demonstrations, or sketches. Such interactions can be more efficient and effective at conveying knowledge than language in many cases. If there are many similar objects in sight, it can be more efficient to point to one than to try and uniquely identify it through language. When trying to teach the agent how to do a new task, it can be more efficient to demonstrate the task being done than to describe it. In some cases, it can be more efficient to sketch a desired configuration of objects than to describe it spatially. Thus, there are additional modalities of interaction that the agent should be able to interpret. The agent should integrate information from these interactions with the dialog context.

Rosie cannot interpret visual gestures, but it can interpret the teacher selecting an object on the computer interface and referring to it using the word 'this' (e.g., *This object is red*). As mentioned above, this information is used to resolve the reference. Rosie cannot interpret sketches or demonstrations of actions or tasks, but can extract information about goal states from demonstrations. For tasks, after the instructor has led the agent through the steps of the tasks, Rosie can infer what the goal of the task is by identifying what the steps accomplished. For games, Rosie can extract a representation of the goal from a demonstration of the goal state instead of a natural language description (e.g., in Tower of Hanoi, it is easier to show the final configuration than to describe it). Figure 1 shows that incorporating goal demonstrations reduces the total number of words needed to completely teach various games and puzzles. This additional modality of interaction allows for more flexible and efficient dialog.

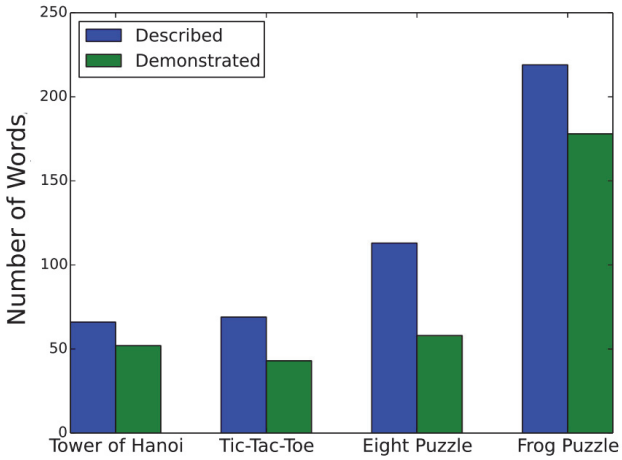

Figure 1: Number of words needed to teach four games through instruction using goal descriptions and demonstrations.

Apply reasoning strategies to interaction

Task-oriented dialog is *mixed-initiative*, *flexible*, and *collaborative*. The onus of communication or information exchange is distributed amongst the participants. Participants advance the dialog in accordance with their intentions and goals and comply with each other's requests. An example of expert-novice dialog from Grosz and Sidner (1986) is shown below.

> E: First you have to remove the flywheel. (expert initiative)
> A: How do I remove the flywheel? (novice initiative)
> E: First, loosen the two Allen head setscrews holding it to the shaft, then pull it off. (expert initiative)
> A: OK. I can only find one screw. Where's the other one? (novice initiative)

In this example, the expert takes the initiative in thinking about the goals of the joint task and proposes a strategy to execute it. Although the novice does not know the structure of the task, she plays an active role in the conversation. She reasons about her environment and knowledge, identifies issues that impede task execution, and guides the interaction to elicit that information from the expert. This frees up the expert from having to closely observe and model the novice in order to provide relevant information.

Often in human controlled interactive learning, such as learning by demonstration, the onus of learning is completely on the human user. The human has to provide good demonstration traces that will result in learning at appropriate levels on generality. However, Rosie is designed to be active and plays a useful role in knowledge acquisition.

Whenever it is unable to make progress on the task, either because it does not understand the words in the instruction, it does not know the task goals, or it does not know which action to take next, it changes the state of interaction by asking a relevant question. Further, it does not completely rely on the human instructor to provide complete instruction, but instead uses its knowledge of the domain to explore possible task executions. For example, once given the goal of a task, Rosie attempts to generate a plan through an internal search of the action space using its internal models of the primitive actions. It asks for help only if the solution is beyond its search-depth horizon. Thus, the human instructor can rely on Rosie avoid asking questions that have obvious answers and to guide interaction to where it needs to make progress on the task.

Figure 2 shows how the number of interactions with the instructor (left y axis) decreases (from left to right) as the agent relies more on its own internal planning when learning how to execute a task. However, this comes at the cost of exponential increases in processing time (as measured in decision cycles).

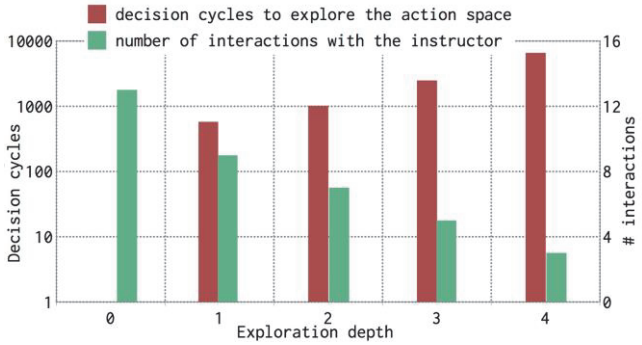

Figure 2: Integrating information communicated through task-oriented dialog with agent planning for efficient interaction.

Integrate prior knowledge

During long-lived human robot interactions, a robot will be commanded to do different tasks, possibly in multiple environments. Interacting with the robot to make it perform these tasks may involve giving it a sequence of actions to perform or a set of conditions to achieve or attend to. An agent that does not learn from these interactions will require these commands to be repeated when given the same or similar task. In human conversations, references are often made to previous interactions, using a label that refers to an abstraction over multiple concepts. Humans often structure teaching interactions to hierarchically build up these structures in order to facilitate accessibility (teaching everything at the most primitive level is tedious) and transference of knowledge (each concept in a hierarchy can potentially transfer). This is a fundamental capability necessary for the agent to be able to extend the types of interactions it supports and increase the efficiency of communication.

In Rosie, there are multiple ways in which the agent integrates prior knowledge and takes advantage of hierarchical teaching strategies. Rather than teaching a complex action, such as *cook*, by using only primitive actions (*pick up, put down, ...*), the task can be decomposed into intermediate actions that are easier to teach, such *place*. In the future, these actions can be used by name to teach more complex actions, such as *serve*, which uses both *cook* and *place* (Mohan & Laird, 2014). Beyond increasing the efficiency of subsequent commands, it provides a higher level of abstraction, which is more accessible for the teacher.

Any concept that is taught, including visual properties, spatial relations, procedural actions, and even specific goals and actions for games, can be used in subsequent interactions by name, which for now we assume is unique. Therefore, Rosie transfers knowledge between games with similar actions and goals, such as *Tic-Tac-Toe* and *Three Men's Morris,* which both have the goal of *Three-in-a-row*. Figure 3 shows the number of interactions that are used to teach three games, separately and sequentially (left to right *Connect 3* followed by *Tic-Tac-Toe* and *4 Queens*). In sequential teaching, the agent demonstrates the benefit in efficiency from the transfer of knowledge between games. The definitions of these games share not only goals and actions, but also spatial relations, actions, and visual properties (Kirk & Laird, 2014), so that the number of interactions required to teach *Tic-Tac-Toe* after learning *Connect 3* is significantly less than when teaching *Tic-Tac-Toe* from scratch because of the transfer of concept learned in *Connect 3*. There is a similar result for *4 Queens*, but it is not as dramatic because *4 Queens* does not share as many concepts with the other two tasks.

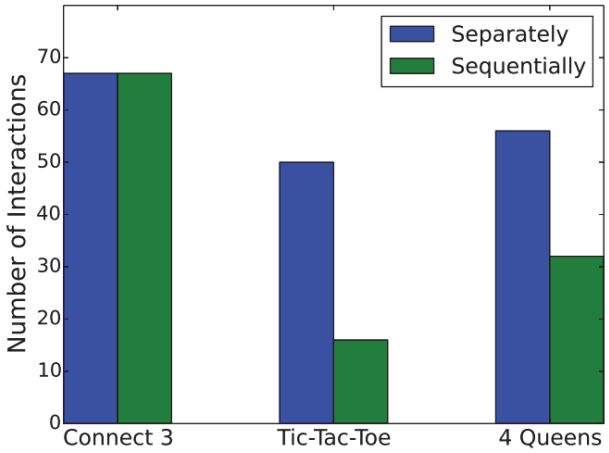

Figure 3: Transfer of knowledge of concepts evaluated by number of interactions needed to describe games separately and sequentially (left to right).

Recent work has extended this capability to include teaching abstractions for any collection of state features,

such that custom features for an environment or task, such as a *protected* location for a specific board game, can be learned and is available for teaching subsequent concepts.

Implement a model of a collaborator's knowledge, intentions, and goals

When working together on a task, interactions between collaborators are *goal-oriented*. Participants communicate with each other in pursuit of goals, which may pertain to manipulating and navigating their environments, changing each other's belief states, gathering information about the task and environment, or comprehending what is being said. A key component of reasoning about interactions is knowing what the collaborator perceives or knows about the task and the environment, what their goals are, and what they intend with their communication. This is useful in generating the right responses to utterances. In a teaching scenario, the intention behind the question *where is Golden Gate Park?* is to test the learner's knowledge. *San Francisco* may be considered a correct answer here. However, in a navigation scenario the same question may be asked to get to a more precise location or to get directions.

Rosie does not explicitly model or reason about the intent and goals of its instructor. Some of this knowledge is implicitly encoded in rules that support the information exchange protocol. Questions from the instructor are always assumed to be tests of knowledge, imperative sentences as action commands that should be executed in the environment. Similarly, Rosie assumes that on asking questions, an informative response will be provided. These assumptions have been useful in the ITL domain; however, they will not apply when there are several instructors who have different goals or when the instructor does not know the task as well. As we extend Rosie to not only function as a learner but to also participate in collaborative task execution or teach another agent tasks, it will need to explicitly model and reason about its collaborators' state.

Communicate the agent's knowledge, intentions, and goals

For a human to effectively collaborate with a robotic agent, it is important that the human can build up a model of the internal state of the agent. When communicating knowledge to the agent, it is useful to judge how effective the interaction has been and how much the agent has learned. This allows the human to tailor their interactions to better fit the capabilities of the agent. The human may correct incorrect knowledge or choose to communicate it in a different way. Thus the agent must be able to describe its own knowledge to the human. If the agent also has significant autonomy, the human may wish to know the current goals and intentions of the agent to better understand what the agent is currently doing and what it is trying to achieve. Thus the agent must be able to describe its current goals and intentions and explain why it believes its current actions will lead to the goal.

We have taken some small steps towards meeting this requirement. When teaching Rosie visual properties and spatial relations, the instructor can ask the agent questions to judge what it has learned. Some examples include *What is this?, What is in the pantry?, Is the red block on the table?,* and *What color is this?* Rosie currently cannot communicate any knowledge about tasks or games (e.g., describe a valid move in Tic-Tac-Toe or describe the goal of storing a block), and cannot answer questions about its own goals. Currently the latter is not very useful because Rosie only performs actions in response to the instructor's commands. One benefit of our system is that most of the agent's knowledge and goals are encoded as declarative representations in either working memory or semantic memory, making them easily accessible to the agent. For efficiency, Rosie dynamically compiles its interpretation of declarative knowledge into rules, so that in many cases it has a fast procedural representation of knowledge in addition to an accessible declarative representation.

Support informative failure states

Human interactions do not always go smoothly, with communications often being ambiguous or incomplete. However, humans can reason about comprehension failures and direct the ongoing interaction with their collaborators to address those failures. They ask for repetitions, clarifications, or explanations. Such behavior requires that the agent maintain information about why a failure occurred and what information is required to address that failure.

Rosie has some capability for handling failures. Its parsing is robust to minor ungrammatical lapses, and when it is unable to parse or semantically interpret a sentence, it will ask for a rephrasing. When novel words are used in a command, including new adjectives, prepositions, or verbs, it will ask for definitions, and reparse the original sentence with what it has learned. This allows the human to give commands without knowing precisely what concepts Rosie already knows.

Conclusion

Although the implementation of Rosie is only a single case study, the analysis should be informative for similar HRI projects, especially robots that engage in task-oriented dialog. A robot attempting to collaborate on a task that lacks these identified requirements would demonstrate clear deficiencies, especially along the criteria we defined: generality, efficiency, and effectiveness. An agent that does not support informative failure states cannot give meaningful information to assist the human in making quick corrections or adjustments. If the agent doesn't incorporate context for interpretation, interactions that are not ambiguous may appear to be ambiguous, requiring further interactions to clarify their meaning. Failure of the agent to apply rea-

soning knowledge will force the human to explain with additional interactions, even when the answer seems obvious. Communication of particular concepts may be more efficient in different interaction modalities. An agent that does not integrate prior knowledge cannot learn higher-level abstractions to support more efficient interactions over long-lived experiments.

General, efficient, and effective task-oriented dialog requires the agent to apply a large variety of information about the environment, shared context of dialog and events, and current state of knowledge. For agents like Rosie that are interactively learning tasks, online knowledge acquisition also requires effective communication. This circular dependency is a strength not a weakness, where additional knowledge can improve the quality of interactions, and better interactions can facilitate the process of adding more knowledge.

There are likely many other requirements that will be exposed by exploring more tasks, domains, and agent architectures. Many HRI problems have not been explored in our work, such as modeling emotional states of humans or using other communication modalities like visual gestures or teleoperation. Furthermore, the accessibility of communication is an important criterion we have not been able to evaluate thoroughly. Researchers that know intimately the details of the agent architecture, the representation of the environment, and supported types of interactions make poor test subjects. We need more experiments on observing human-robot task-oriented dialog to evaluate how easy it is for an average person to communicate. Human-human trials on similar tasks would be informative as to the exact forms of interaction and strategies that occur naturally.

References

Bangalore, S., Di Fabbrizio, G., & Stent, A. (2008). Learning the Structure of Task-Driven Human-Human Dialogs. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(7), 1249–1259.

Engelmore, R., and Morgan, A. eds. 1986. *Blackboard Systems*. Reading, Mass.: Addison-Wesley.

Grosz, B. & Sidner, C. 986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics*, 12(3), 175–204.

Kirk, J. R. & Laird, J. E. (2014) Interactive Task Learning for Simple Games. *Advances in Cognitive Systems*, vol. 3, pp. 13–30.

Kollar, T., Tellex, S., Roy, D., & Roy, N. (2010). Toward Understanding Natural Language Directions. In Proceeding of the Fifth *ACM/IEEE International Conference on Humanrobot Interaction*, pp. 259–267. Osaka, Japan: IEEE Xplore.

Laird, J. (2012). The Soar Cognitive Architecture. Cambridge, MA: MIT Press.

Laird, J. (2014). Report on the NSF-funded Workshop on Taskability (Interactive Task Learning).

Litman, D. & Allen, J. (1987). A Plan Recognition Model for Subdialogues in Conversations. *Cognitive Science*, 11(2), 163–200.

Litman, D. & Allen, J. (1990). Discourse Processing and Commonsense Plans. *Intentions in Communication*, 365–388.

Matuszek, C., Fitzgerald, N., Zettlemoyer, L., Bo, L., & Fox, D. (2012). A Joint Model of Language and Perception for Grounded Attribute Learning. *Proceedings of the Twenty Ninth International Conference on Machine Learning*. pp. 1671–1678.

Mohan, S. (2015). From Verbs to Tasks: An Integrated Account of Learning Tasks from Situated Interactive Instruction. Ph.D. dissertation, University of Michigan, Ann Arbor.

Mohan, S. & Laird, J. E. (2014). Learning Goal-oriented Hierarchical Tasks from Situated Interactive Instruction. *Proceedings of the Twenty Eighth AAAI conference on Artificial Intelligence*. AAAI Press.

Mohan, S., Mininger, A., Kirk, J. R., & Laird, J. E. (2012). Acquiring Grounded Representations of Words with Situated Interactive Instruction. *Advances in Cognitive Systems*, vol. 2, pp. 113–130.

Oviatt, S. L. & Cohen, P. R. (1991). Discourse Structure and Performance Efficiency in Interactive and Non-Interactive Spoken Modalities. *Computer Speech & Language*, 5(4), 297–326.

Scheutz, M., Cantrell, R., & Schermerhorn, P. (2011). Toward Human-Like Task-based Dialogue Processing for HRI. *AI Magazine*.

Tellex, S., Kollar, T., Dickerson, S., & Walter, M. (2011). Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation. *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*. San Francisco, CA: AAAI Press.