Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy*

Dennis Perzanowski, Alan C. Schultz, William Adams, and Elaine Marsh

Navy Center for Applied Research in Artificial Intelligence
Naval Research Laboratory
Codes 5512 and 5514
Washington, DC 20375-5337
< dennisp | schultz | marsh | adams > @aic.nrl.navy.mil

Abstract

Intelligent mobile robots that interact with humans must exhibit adjustable autonomy; that is, the ability to dynamically adjust the level of self-sufficiency of an agent depending on the situation. When intelligent robots require *close* interactions with humans, they will require modes of communication that enhance the ability for humans to communicate naturally and that allow greater interaction, as well as adapt as a team member or sole agent in achieving various goals. Our previous work examined the use of multiple modes of communication, specifically natural language and gestures, to disambiguate the communication between a human and a robot. In this paper, we propose using *context predicate*s to keep track of various goals during human-robot interactions. These context predicates allow the robot to maintain multiple goals, each with possibly different levels of required autonomy. They permit direct human interruption of the robot, while allowing the robot to smoothly return to a high level of autonomy.

Introduction

The tasks and goals of the robotic system we have been developing require tight human and robot interactions. Combined human/robot systems employing cooperative interaction to achieve those tasks require that goals and motivations originate either from the human or from the robot. It may be necessary for either of these agents (the human or the robot) to assume the responsibility of instantiating goals which direct the combined human/robot team towards completion of its task. We refer to systems with this property as *mixed-initiative systems*, i.e. the initiative to dictate the current objective of the system can come from the robot itself or from a human.

* This work is funded in part by the Naval Research Laboratory and the Office of Naval Research.

autonomy is a critical requirement. Adjustable autonomy allows systems to operate with dynamically varying levels of independence, intelligence, and control. In these systems, a human user, the robot, or another robot, may adjust each team member's "level of autonomy" as required by the current situation. This may be done by fiat, but most frequently in human situations, adjustments are made cooperatively, swiftly, and efficiently. Our research addresses the case of human-robot interactions, where human interaction with the robot will require the robot to smoothly and robustly change its level of autonomy. Further, we believe that a clue to how systems can adjust their autonomy cooperatively is by keeping track of the goals of a task or mission and then acting on an immediate goal as it relates to that agent's role in completing the mission.

The need for adjustable autonomy is clear in situations where intelligent mobile robots must interact with humans.

Consider the following examples:

Several dozen micro air vehicles are launched by a Marine. These vehicles will have a mission to perform, but depending on the unfolding mission, some or all of the vehicles may need to be redirected on the fly, at different times, and then be autonomous again.

Groups of autonomous underwater vehicles involved in salvage or rescue operations may start by autonomously searching an area, but then need to be interrupted by a human or another robot to be redirected to specific tasks.

A planetary rover interacts with human scientists. Because of the communication time lag in this situation,

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1999		**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Goal Tracking in a Natural Language Interface: Towards Achieving Adjustable Autonomy	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Navy Center for Applied Research in Artificial Intelligence,Naval Research Laboratory,Code 5510,Washington,DC,20375**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Intelligent mobile robots that interact with humans must exhibit adjustable autonomy; that is, the ability to dynamically adjust the level of self-sufficiency of an agent depending on the situation. When intelligent robots require close interactions with humans, they will require modes of communication that enhance the ability for humans to communicate naturally and that allow greater interaction, as well as adapt as a team member or sole agent in achieving various goals. Our previous work examined the use of multiple modes of communication, specifically natural language and gestures, to disambiguate the communication between a human and a robot. In this paper, we propose using context predicates to keep track of various goals during human-robot interactions. These context predicates allow the robot to maintain multiple goals, each with possibly different levels of required autonomy. They permit direct human interruption of the robot, while allowing the robot to smoothly return to a high level of autonomy.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**6**	

autonomy is critical to the safety of the vehicle. However, the human must be able to exert lower levels of control to perform various experiments.

In such tasks, for example, humans will be exerting control over one or more robots. At times, the robots may be acting with full autonomy. However, situations will arise where the human must take low-level control of individual robots for short periods, or take intermediate level of control over groups of robots by giving them a new short-term goal which overrides their current task. The robots must be able to smoothly transition between these different modes of operation.

In similar situations, where intelligent mobile robots must interact closely with humans, *close* interaction and natural modes of communication, such as speech and gestures, will be required. Our previous work (Perzanowski, Schultz, and Adams, 1998) examined the use of such modes of communication in order to disambiguate the speech input. However, in situations where agents must interact with other agents, human and/or robotic, this capability must be coupled with an awareness of the status of mission goals: those achieved, sub-goals—perhaps previously unknown—needing completion, and where the agent is in the overall mission.

Currently, we are exploring the use of *context predicates* to keep track of various goals during human-robot interactions. These context predicates allow the robot to track the status of a goal, and even maintain multiple goals, each with possibly different levels of required autonomy. They permit direct human interaction when necessary, and allow the robot to smoothly return to a high level of autonomy.

In the following paper, we will describe the robot platform and supporting software. Next, we will briefly describe our previous work on multi-modal communication. We will discuss how we presently use context predicates to track goal status and goal attainment and discuss some related and future work. We will conclude with some general thoughts on how our current work can be applied to achieving adjustable autonomy.

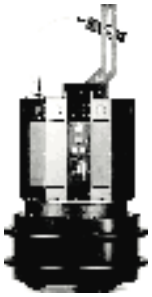

Figure 1: A Nomad 200 mobile robot with mounted camera

Robotic Platform

For our research in developing a natural language and gestural interface to a mobile robot we have been employing a Nomad 200 robot (see Figure 1), equipped with 16 Polaroid sonars and 16 active infrared sensors.

Gestures are detected with a structured light rangefinder emitting a horizontal plane of laser light. A camera mounted above the laser is fitted with a filter tuned to the laser frequency. The camera observes the intersection of the laserlight with any objects in the room, and the bright pixels in the camera's image are mapped to XY coordinates. When the sensors on the robot detect a vector or a line segment within the limitations of its light striping sensor, and a command is sent to move in some direction or a specified (gestured) distance, a query is made of the gesture process on the robot to see if some gesture has been perceived. Whether or not a particular command requires a specific gesture is determined, and appropriate commands are sent to elicit an appropriate response. The mapping of the speech input and the perceived gesture is a function of the appropriateness or inappropriateness and the presence or absence of a gesture during the speech input. The two inputs, the semantic interpretation mapped into a command interpretation and the gesture signal, are then translated to a message, which is then sent to the robot in order to produce an appropriate action or reaction. A more detailed analysis of how the system processes the visual cues, namely the gestures, in conjunction with the natural language input, can be found in (Perzanowski, Schultz, and Adams, 1998).

Multi-Modal Communication

The first stage of our interface was built relying on the interaction between natural language and gesture to disambiguate commands and to provide complete information where one of the two channels of communication lacked some specific or required information. Thus, for example, the utterance, “Go over there” may be perfectly understandable as a human utterance, but in the real world, it does not mean anything if the utterance is not accompanied by some gesture to indicate the locative goal.

For this work, we assumed that humans frequently and naturally use both natural language and gesture as a basis for communicating certain types of commands, specifically those involved in issuing directions. When a human wishes to direct a mobile robot to a new location, it seemed perfectly natural to us to allow the human the option of using natural language or natural language combined with gesture, whichever was appropriate and produced a completely interpretable representation which could then be acted upon.

Coincidentally, we did not incorporate any hardware devices, such as gloves (McMillan 1998), for inputting gesture information. In order to keep our interactions as “natural” as possible, we have not included such devices

which would, in some sense, restrict the human in interacting with the robot.

Furthermore, we did not permit gestures in isolation because we believed that their use took the communicative act out of the natural realm and made it a more symbolic act, which we did not wish to pursue at that point. We are not, however, ruling out isolated, symbolic gestures or symbolic gestures in combination with speech as possible means of efficient interaction with mobile robotic systems. We simply leave their consideration for future work.

Just as others, such as (Konolige and Myers 1998), have incorporated gesture recognition as part of the attention process in human-robot interactions, we have concentrated on the naturalness of the gesture, along with its ability to disambiguate natural language. However, we restricted the types of communication in this interface to a model of communication characterized as a *push* mode (Edmonds 1998). By this, we mean the human basically provides all the input, and the mobile robot merely acts as a passive agent, reacting only to those commands issued by the human participant.

Input was restricted to commands that involved achieving only one goal. If any interruptions occurred, or if intervening goals made it necessary for the primary directive to be kept on hold, the system failed. These were obvious limitations of the system and on the naturalness of the interaction, but it was the first step toward integrating natural language and gesture in an interface to a mobile robot.

Autonomy was simply not an issue for this system, since the robot could only react to the commands issued by the human. However, once we began to parse fragmentary verbal input or incomplete sentences, we found that we also had a mechanism for tracking and determining the status of achieved and unachieved goals. Questions of autonomy began surfacing. Therefore, to see one way in which adjustable autonomy can be achieved, we will focus on how our system analyzes the natural language input. The gestural input will also be considered whenever it becomes crucial for the interpretation of a command.

Our analysis of the natural language input requires a full syntactic parse of the speech input. We do not employ a stochastic or probabilistic parsing technique (Charniak, 1997) since we believe our corpus is too small at this time to make this technique efficient. Given a full syntactic parse, a semantic interpretation is obtained, utilizing our in-house natural language processing system (Wauchope 1994). When a complete representation is obtained, it is then translated to an appropriate message that the robotic system can process and an action is performed.

A Brief Overview of System Capabilities

We turn now to a short example of how the earlier version of our interface processed commands. This functionality remains in the current version of the interface.

If a human wants the robot to move to a new or different location, the human can either utter a sentence, such as one of the sample set of sentences in (1), or the human can utter a sentence along with performing an appropriate gesture.

(1) (a) Go to the left/right.
(b) Move to the right/this way.
(c) Back up this far/10 inches.
(d) Go to waypoint one/the waypoint over there.

Thus, for example, if (1a) or (1b) are uttered while the human correctly points in the direction corresponding to the robot's right, left, or in a specific direction, the robot responds appropriately by moving in the desired direction. If an inappropriate gesture is made, the system responds with an error message, giving the human some notion of whether a contradictory gesture was given. If no gesture is made but one is needed, as in (1b,c, and d), the robot complains about the incompleteness of the command. These responses usually consist of canned messages, such as "I'm sorry. You told me to go one way, but pointed in another direction. What do you want me to do?"

We now turn to our proposal to use context predicates to enhance the system's capabilities to track its goals, thereby introducing a capability to provide greater autonomy in human-robot interactions.

Using Context Predicates

As a first step in our attempt to provide greater autonomy in robotic control, the natural language and gestural interface was enhanced to enable the processing of incomplete and/or fragmentary commands during human-robot interactions. (2) gives an example of a small dialog containing a fragmentary command (2c).

(2) (a) Participant I: Go over there. [no gesture accompanies verbal input]
(b) Participant II: Where?
(c) Participant I: Over there. [gesture accompanies verbal input]

(2c) is a fragment because an entire command containing a verb is not given (see (2a)). Linguistically, (2c) consists only of the adverbial expression of location "over there". The system must somehow remember that the correct action to take is found in the verb of a preceding sentence, namely "go" of (2a).

On a very basic level, this ability to go back and pick out an appropriate action for a fragment currently being processed requires that certain kinds of information be stored for later use. To achieve this functionality, we create a stack of predicates, or verbs and their essential arguments, at the beginning of an interaction, and continually update it during the interaction. We call this stack the *context predicates*. If it becomes necessary to obtain information at a later time in the human-robot interaction, the information is available. For the processing of sentence fragments, it is a simple matter of obtaining the

correct verb or action to go with the fragment by searching the stack in particular ways to be discussed below.

For example, in processing the sentences of (2), a context predicate stack is created with Participant I's utterance (2a). At this point the stack consists of one item, a list that looks something like (3).

(3) ((imper #v5414 (:class move-distance)
(:agent (pron n2 (:class system) you))
(:goal (name n1 (:class loc) there)))
0)

The list contains the action requested, namely "to go" which belongs to a semantic class of verbs we call "move-distance" verbs. We semantically classify verbs in order to make linguistic generalizations and future processing easier. The list also contains one of the arguments of the verb, namely the agent, which in an imperative "imper" sentence is always "you." This pronoun belongs to a semantic class of objects we call "system" nouns or pronouns. These function as agents, and move-distance verbs require agents that are sub-classified as systems. This analysis is part of the semantic component of the natural language understanding system.

The second argument of move-distance verbs is a goal, which in this sentence is the adverb of *loc*ation "there." (Trivially, the word "over" is not included here.) Finally, the digit 0 is incorporated in the list to indicate that the goal has not been completed; i.e., no robotic action has occurred. Identifying numbers for the parts of speech are also provided for later processing and referencing.

Currently, we do not process the robot's responses, such as (2b). However, upon hearing (2b), Participant II's request for more information, the human issues the fragmentary command (2c). It is parsed and its representation looks something like (4).

(4) ((imper #v5415 (:class dummy-verb)
(:agent (pron (:class system) you))
(:goal (name (:class loc) there)))
0)

A stack is now created consisting of the two lists (3) and (4).

The natural language understanding system notes that the verb of the sentence belongs to a class of "dummy" verbs and notes that the goal has not been achieved (0). It requires that these verbal elements, like pronouns, must have an anaphor somewhere in the previous discourse. It looks at the stack and sees that there is a verb belonging to the move-distance class of verbs that also has the same set of arguments; i.e. the goals and agents in both (3) and (4) are identical and its goal has not been achieved as evidenced by the digit 0 in the list of (3). The system concludes that the dummy-verb in (4) must be of the same class as the one in (3). It therefore substitutes the verbal class in (3) into the dummy-verb slot of (4). Given that an appropriate gesture has also been noted during the processing of this command, a message is passed to the robot for appropriate action. The stack is then updated so that all actions involved in this interchange are noted as complete by updating the digit 0 to 1. The use of the digits 0 or 1 simply allows the system to determine whether or not an action has been completed (goal attained). Furthermore, when a conversation becomes lengthier, it is still a simple matter of checking the stack to see which actions have or have not been completed in the stack. The last slot in the representation of the various utterances is somewhat like a record of the context, whether or not some action or predicate has been completed, hence the name *context predicates*.

One might argue that the stack might become too lengthy to handle; however, we are currently investigating ways to keep the stack tractable by incorporating such discourse elements as topic (Grosz and Sidner, 1986) and attentional or focus states (Stent et al., 1999) to dictate how far or deeply into a stack the natural language understanding system should dig for information.

Schematically, we can represent how context predicates are obtained in Figure 2.

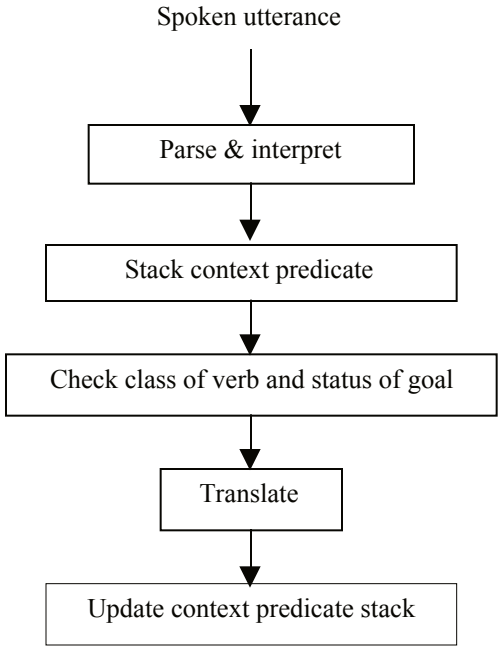

Figure 2: Schematic flowchart for processing utterances

We turn now to a more specific example in which context predicates can be used to track goals.

In this example, let us construct a brief scenario (5).

(5) A human issues a command for a robot partner to go to a particular waypoint by name, such as "Go to waypoint 3." On its way to waypoint 3, the robot confronts some obstacle not previously mentioned by the human that must be moved, such as a box, or opened, such as a closed door, in order to proceed.

Assuming that the robot's planning and navigation components know what to do with obstacles that need to be moved and/or opened, the robot should act independently to figure a way around or through the obstacle and proceed on its mission, which in this case is to proceed to waypoint 3.

Given our system's ability to stack commands as context predicates and to store information about the completion of those actions, (5) can be accomplished with nothing more than the initial command to proceed to a particular location.

At the beginning of the scenario (5), a command is issued and a stack is created consisting of the list (6a).

(6)(a) ((imper #v6600 (:class move-distance)
(:agent (pron n2 (:class system) you))
(:to-loc (null-det n1 (:class waypoint) (:id 3))))
0)

While acting on the verb and its arguments in (6a), the robot encounters the obstacle. The context predicate in (6a) is still marked as unaccomplished. The planning and navigation components independently issue commands for the robot either to move the obstruction or open the closed door. These commands are parsed by the natural language component and their representations and status are stacked along with any other context predicates. Thus, (6b) is added to the stack.

(6)(b) ((imper #v6601 (:class open)
(:agent (pron n2 (:class system) you))
(:patient (noun n2 (:class object) door)))
0)

Once the door is opened, the placeholder for the status of the command is updated, and the context predicate stack is checked for previously uncompleted commands. This stack checking occurs as long as the robot is tasked to do something, and it stops once all of the goals have been attained. Tasking in this context is complete when all context predicates have a final value of 1. So, whenever the stack is revisited and an incomplete predicate is found, the robotic system knows in a sense that a task still needs to be completed and a goal achieved.

In our example scenario the robot's ultimate task is to get to a particular waypoint. Having completed the interrupting task of opening a door, it can now continue on its previous mission, unless of course other interruptions occur, which the planning and navigation components must decide upon and act upon. If actions are taken, their representations are mapped onto the context predicate stack for further comparisons. And so the cycle continues until the first predicate in the stack, move-distance of (6a), receives a value of 1, denoting completion.

This scenario requires that the natural language understanding system and the planning and navigation components onboard the robot can swap information. We are currently implementing this functionality.

Related Work

As we stated previously, we currently do not employ any symbolic gestures (Kortenkamp, Huber, and Bonasso 1996) in our natural language and gestural interface. Presently, all gestures are natural and indicate directions and distances in the immediate vicinity of the two participants of the interaction, namely the human and the robot. In future, however, we intend to incorporate symbolic gestures into the interface and to provide seamless integration of both types of gestures. Later, we hope to permit the user to incorporate symbolic gestures and for the system to know the difference between natural and symbolic communication.

Another mobile robot, Jijo-2 (Matsui et al., 1999), provides natural spoken interaction with an office robot. Natural dialog and a sophisticated vision and auditory system permit Jijo-2 to interact with several humans, to remember conversants and to locate humans to engage in conversation. While our current system does not have such a sophisticated vision or auditory system, we have concentrated on maximizing gestural information from a very limited vision source and developed a natural language component that allows for interrupted and fragmentary dialog. Thus far, our efforts have been constrained by the vision system we have employed, but we believe we have maximized it and shown success in integrating natural language and gesture for interacting with a mobile robot. While a system like Jijo-2 concentrates on natural language and face recognition, for example, we have concentrated on natural language and gesture recognition. We, therefore, have concentrated on developing a natural means of communicating with a mobile robot.

Although we are not claiming that communication with robotic agents must be patterned after human communication, we believe that human/machine interfaces that share some of the characteristics of human-human communication can be friendlier and easier to use. Thus, if a system has vision capability similar to human vision capability, chances are humans will naturally interact with that capability on a machine. The current version of our interface permits a natural way for humans to interact with a mobile robot that has a well-defined but limited vision capability.

Future Work

While we currently employ context predicates to track goals obtained in fragmentary input, we anticipate their use in tracking goals in lengthier dialogs. As a result, greater autonomy will be achieved, since users can expect the robotic system to be able to continue to perform and

accomplish previously stated goals or subsequent logical sub-goals, without the user having to explicitly state or re-state each expected or desired action. The system will be able to engage in immediate actions and commands, as well as obtain previously unattained goals, by utilizing verbal class membership in the context predicates discussed above and noting whether or not predicates within a given context have been completed or not.

For this work, we intend to add another item to the context predicates. We would like to incorporate a kind of prioritization of tasks to determine the order in which actions need to be accomplished when several tasks remain to be completed.

We intend to conduct experiments on the enhanced system in the near future with the intention of incorporating empirical results of those studies for future publication.

Conclusions

Based on our work to develop a natural interface to a mobile robot, we concentrated on natural language and gestures as means of interaction. As we looked at the kinds of communication that humans exhibited during those interactions, we saw that frequently humans use fragmentary or incomplete sentences as input. This led us to incorporate the notion of *context predicates* into the natural language processing module of the interface. Given how context predicates can be a means of tracking goals and their status during human/robot dialogs, we are currently investigating ways to utilize context predicates and goal tracking to permit humans and robots to act more independently of each other. As situations arise, humans may interrupt robot agents in accomplishing previously stated goals. Context predicates allow us to keep track of those goals, whether they have been completed or not, and can even permit a record to be kept of the necessary steps in achieving them. With this capability, the system can return after interruptions to complete those actions, because the system has kept a history of which goals have or have not been achieved. This capability of our system allows both the human and the robot in these interactions to work at varying levels of autonomy when required. Humans are not necessarily required to keep track of robot states. The system does, and the robot is capable of performing goals as they are issued, even if an intervening interruption prevents an immediate satisfaction of that goal.

The incorporation of context predicates to track goals will be a necessary capability to allow adjustable autonomy in robots, which in turn permits the kinds of interactions and communication in the mixed-initiative systems we are developing.

References

Charniak, E. 1997. Statistical Parsing with a Context-free Grammar and Word Statistics. In Proceedings of the Fourteenth National Conference on Artificial Intelligence, 598-603. Menlo Park, CA: AAAI Press/MIT Press.

Edmonds, B. 1998. Modeling Socially Intelligent Agents. *Applied Artificial Intelligence* 12:677-699.

Grosz, B. and Sidner, C. 1986. Attention, Intentions, and the Structure of Discourse. *Computational Linguistics* 12(3):175-204.

Konolige, K. and Myers, K. 1998. The Saphira Architecture for Autonomous Mobile Robots. In Kortenkamp, D., Bonasso, R. P., and Murphy, R. eds. *Artificial Intelligence and Mobile Robots: Case Studies of Successful Robot Systems*, 211-242, Menlo Park, CA: AAAI Press/MIT Press.

Kortenkamp, D., Huber, E., and Bonasso, R.P. 1996. Recognizing and Interpreting Gestures on a Mobile Robot. In Proceedings of the Thirteenth National AAAI Conference on Artificial Intelligence, 915-921. Menlo Park, CA: AAAI Press/MIT Press.

Matsui, T., Asoh, H., Fry, J., Motomura, Y., Asano, F., Kuria, T., Hara I., and Otsu, N. 1999. Integrated Natural Spoken Dialogue System of Jijo-2 Mobile Robot for Office Services. In Proceedings of the Sixteenth National Conference on Artificial Intelligence, 621-627. Menlo Park, CA: AAAI Press/MIT Press.

McMillan, G.R. 1998. The Technology and Applications of Gesture-based Control. In RTO Lecture Series 215: Alternative Control Technologies: Human Factors Issues, 4:1-11. Hull, Québec: Canada Communication Group, Inc.

Perzanowski, D., Schultz, A.C., and Adams, W. 1998. Integrating Natural Language and Gesture in a Robotics Domain. In Proceedings of the IEEE International Symposium on Intelligent Control: ISIC/CIRA/ISAS Joint Conference, 247-252. Gaithersburg, MD: National Institute of Standards and Technology.

Stent, A., Dowding, J., Gawron, J.M., Bratt, E.O., and Moore, R. 1999. The CommandTalk Spoken Dialogue System. In 37th Annual Meeting of the Association for Computational Linguistics: Proceedings of the Conference, 183-190. San Francisco: Association for Computational Linguistics/Morgan Kaufman .

Wauchope, K. 1994. Eucalyptus: Integrating Natural Language Input with a Graphical User Interface, Technical Report, NRL/FR/5510--94-9711, Navy Center for Applied Research in Artificial Intelligence. Washington, DC: Naval Research Laboratory.