# Uncertainty, Utility, and Misunderstanding: A Decision-Theoretic Perspective on Grounding in Conversational Systems

**Tim Paek**

Department of Psychology
Stanford University
Stanford, CA 94305
paek@psych.stanford.edu

**Eric Horvitz**

Adaptive Systems and Interaction
Microsoft Research
Redmond, WA 98052
horvitz@microsoft.com

### Abstract

Uncertainty plays a central role in the handling of misunderstanding in dialog. People engaged in conversation typically take a sequence of actions to establish and maintain mutual understanding—a process referred to as *grounding*. We explore representations and control strategies for grounding utterances founded on performing explicit probabilistic inference about failures in communication. The methods are informed by psychological studies and founded on principles of decision making under uncertainty. We delineate four distinct levels of analysis for representing uncertainty and describe a computational framework for guiding action in an automated conversational system. We demonstrate how the framework captures grounding behavior by facilitating collaborative resolution of uncertainty as implemented in a spoken interactive dialog prototype called the *Bayesian Receptionist*.

## Introduction

Researchers in linguistics, psychology, and artificial intelligence have argued that dialog may be best regarded as a type of *joint activity* (Clark, 1996; Cohen and Levesque, 1994; Grosz and Sidner, 1990; Suchman, 1987). In a joint activity, such as dancing, participants act in coordination with each other. Joint activities are goal-centric and constrained in terms of setting, role, and, above all, on the kinds of allowable and appropriate contributions (Levinson, 1992). Participants derive explicitly or implicitly a common set of beliefs about the activity, and they drive towards mutual understanding of their intentions and actions—a process referred to as *grounding* (Clark & Brennan, 1991; Clark & Schaefer, 1987, 1989; Clark & Wilkes-Gibbs, 1990).

Just as a dance is more than the sum of individual autonomous motions, a conversation is more than a structured sequence of utterances. People engaged in conversation elegantly coordinate the presentation and acceptance of utterances to achieve and confirm mutual understanding. In the process, they frequently must make decisions under uncertainty. Such actions are targeted at reducing the risk of misunderstanding and contributing to the communication goals. Uncertainty is ubiquitous in dialog. As examples, a listener may be uncertain about the articulation of an utterance and a speaker may be uncertain about the attentiveness or comprehension of the listener. Although participants may tolerate a small degree of uncertainty, an excessive amount in a given context can lead to misunderstanding with all of its associated costs, including the premature termination of the joint activity and, ultimately, poor world actions.

We have been pursuing within the *Conversational Architectures* project automated decision making methods that exploit representations at multiple levels of uncertainty about the ongoing status of a dialog. We believe that such methods promise to change in a qualitative manner the experience and effectiveness of communicating with a computational system. In this paper, we delineate four levels of analysis and describe representations, control structures, and inference strategies for managing uncertainty within and between levels. We illustrate the operation of the methods by reviewing sample interactions drawn from conversation between a user and a prototype system named the *Bayesian Receptionist* (Horvitz & Paek, 1999). The *Bayesian Receptionist* employs Bayesian inference, natural language parsing, and speech recognition to guide dialog about tasks typically handled by front desk receptionists at the Microsoft corporate campus. We also discuss the role observational and psychological studies played in shaping the structure and setting the parameters of the system.

## Toward Robust Automated Grounding

In human–computer dialog, the success of spoken interactive systems relies critically upon the abilities of a set of component technologies, including speech recognition, user modeling, text-to-speech, and natural language processing. To date, conversational systems built as assemblies of such components are fraught with multiple classes of failure. Failures such as speech

*AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems*, Cape Cod, MA. November 5-7, 1999. 

recognition errors are simply thrown into a conversation and users are left with the task of deciphering them. Users are put in the position of relying on their conjectures about the nature of the specific failure and functioning of the overall architecture of the system. In stark contrast to the typical fragility of automated conversational systems, people manage quite well in the midst of uncertainties and imprecision that characterize human–human dialog. They display not only the ability to reason about key uncertainties and their costs, but also exploit strategies such as grounding for collaboratively resolving them.

The computational framework we develop in this paper is motivated by the idea that a dialog system should be able to maintain a conversation without having the luxury of perfect speech recognition, language understanding, or precise models of the goals and attention of users. Such methods should be based on representations and inference machinery that can provide awareness and vigilance about potential failures by considering the uncertainties about possible sources of misunderstanding (Horvitz & Paek, 1999). Interactions with such a system would proceed much in the way that a conversation might with interlocutors of impaired hearing or deficient language skills. Just as people employ communicative strategies like grounding to compensate for their impaired abilities, a dialog system should respond similarly, weighing the costs and benefits of taking particular kinds of compensatory measures.

Our framework is built upon earlier research that has investigated how people collaboratively contribute to a conversation at successive levels of mutual understanding through grounding (Clark & Schaefer, 1987, 1989). While researchers have examined the relationship of these multiple levels with miscommunication (Brennan & Hulteen, 1995; Dillenbourg et al., 1996; Traum, 1994; Traum & Dillenbourg, 1996, 1998), relatively little work has focused on exploiting uncertainty; for example, by explicitly quantifying uncertainty in terms of probabilities at each level. The framework we present broadens the scope of previous models of grounding and referential expressions (Edmonds, 1993; Heeman, 1991; Heeman & Hirst, 1992; Hirst et al., 1994) by highlighting the efficacy of Bayesian networks and decision theory to reason about uncertainty before and during misunderstanding. Furthermore, the introduction of decision theory allows systems to use expected utility to provide fine-grain, context-sensitive guidance of compensatory measures, rather than relying solely on ad hoc procedures (Brennan, 1998).

We adapt the psychological theory of dialog as a joint activity as the central basis of human–computer interaction. We exploit data from psychological studies to shape the structure of Bayesian networks, set parameters, and consider actions in accordance with utilities elicited from users.

We now delineate the distinct levels of analysis and describe the value of explicitly representing and manipulating information about these levels. Next, we introduce a decision-making framework for grounding utterances. Finally, we show examples of inference working in concert at multiple levels of uncertainty in the *Bayesian Receptionist*.

## Four Levels of Representation and Analysis

Previous attempts to model dialog as a joint activity have focused primarily on the coordination of communication based on propositional beliefs (Cohen & Levesque, 1991; see Haddadi, 1995 for a review). This work has led to significant strides in the logical formalization of joint intentions. However, as researchers in grounding have noted, such work has overlooked critical aspects of joint coordination in dialog that span several different levels of mutual understanding. For example, speakers often repeat themselves if they believe they were not heard since, in a joint activity, it is not enough to just produce utterances; speakers must check that their utterances were attended to and that listeners are still engaged in the activity at hand.

Taking inspiration from Clark (1996), speakers and listeners ground mutual understanding at four levels, as shown in Figure 1.

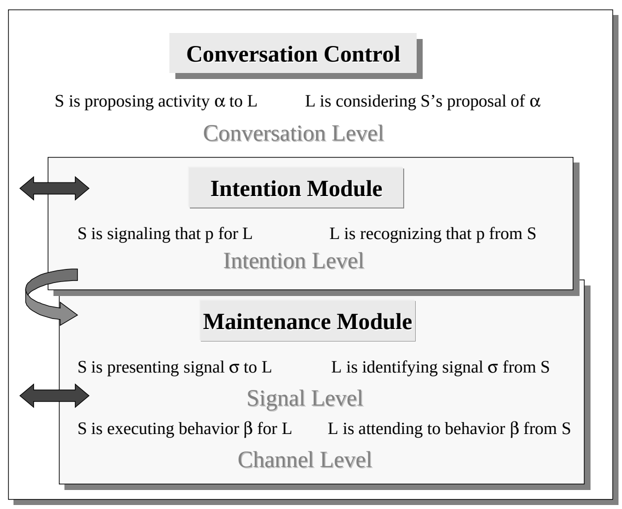


Figure 1. Four levels of representation for inference and decision making under uncertainty.

At the most basic level, which we denote as the *channel level*, a speaker *S* attempts to open a channel of communication by executing behavior $\beta$, such as an utterance or action, for listener *L*. However, *S* cannot get *L* to perceive $\beta$ without coordination: *L* must be attending to and perceiving $\beta$ *precisely* as *S* is executing it.

At the next higher level, the *signal level*, *S* presents $\beta$ as a signal $\sigma$ to *L*. Not all behaviors are meant to be signals, as for instance, the behavior of a listener scratching an itch during a conversation is irrelevant to the content of the interaction. Hence, *S* and *L* must coordinate on what *S* presents with what *L* identifies.

The *intention level* is where interlocutors construe the semantic content of signals. To date, research on conversational systems has been focused almost entirely on the intention level. At this level, *S* signals some proposition *p* for *L*. What *L* recognizes to be the goal of *S* in signaling $\sigma$ is *how L* will arrive at *p*. Note that the signal $\sigma$ is different from the *goal* of *S* in using $\sigma$ (*e.g.*, in indirect speech acts). By focusing on the goals of *S*, the intention level treats the "speaker's meaning" (Grice, 1957) as primary. *S* cannot convey *p* through $\sigma$ without *L recognizing* that *S* intends to use $\sigma$. This again takes coordination.

Finally, at the *conversation level*, *S* proposes some joint activity $\alpha$ which *L* considers and takes up. A proposal solicits an expected response defined by $\alpha$. For example, in an indirect speech act such as "I have to go to the North Campus," meaning "Please call a shuttle to the North Campus," *S* is proposing an activity for *S* and *L* to carry out jointly—namely, that *S* gets *L* to call a shuttle. *S* cannot get *L* to engage in the activity without the coordinated participation of *L* in calling a shuttle.

In short, all four levels require coordination and collaboration in order to achieve mutual understanding.

### Uncertainty and Costs

Actions at all four levels are typically taken under uncertainty. For example, at the lowest level, interlocutors may be uncertain about channel fidelity; hence, they may search for physical signs of attentiveness such as eye gaze or head nodding. Indeed, listeners often display evidence of understanding through backchannels such as "uh huh" (Goodwin, 1986).

Distinct actions under uncertainty are often associated with different costs. For example, if listeners are uncertain about channel fidelity, the cost of asking for a repeat may be slight compared to the cost of continuing along without grounding the recent utterances. With confirmations such as "Did you get that?," the benefits of improving the likelihood of mutual understanding may outweigh the cost of asking, considering the larger expected cost associated with the potential need to retrace later on. Costs and uncertainty also vary according to the media of interaction (Clark & Brennan, 1991).

A spoken dialog system needs to explicitly represent plausible actions, uncertainties, and costs if it is to recover from misunderstanding in a robust and natural manner. Previous approaches centering on the use of propositional logic to model joint action may work well for activities associated with deterministic goals. However, logical methods do not have the representational richness to support in an efficient manner the common situation where goals and their instantiations are uncertain. Furthermore, an analysis of goals at the intention level may fail due to problems at lower levels. Misunderstanding can cross *all* levels of mutual understanding. Rather than disregarding uncertainty, a dialog system should heed it in the way that humans appear to do, weighing the costs and benefits of taking particular actions in light of the uncertainty.

## A Decision-Making Framework for Dialog

Unlike previous models of grounding that represent multiple levels, a decision-making framework allows for uncertainty about what level of misunderstanding a system may be encountering. Rather than having a problem at just one level, and taking action for that level only, a system may be uncertain about which level to investigate, as well as what the costs and benefits of exploring different actions at different levels may be. This problem is especially serious for modular dialog systems that integrate information from a wide variety of component technologies.

We employ Bayesian reasoning and expected value decision making to identify ideal actions in dialog, taking into consideration uncertainties about communication fidelity and meaning, and the potentially varying costs and benefits of alternate actions taken under these uncertainties. We compute the likelihood of states of interest that we cannot observe directly with Bayesian networks. Bayesian networks have been used previously in several user modeling projects (*e.g.,* see Conati et al., 1997; Horvitz, 1997; Horvitz et al., 1998). Before detailing the representations and inference strategies we employ for resolving uncertainty at the four levels, we briefly discuss the control infrastructure of the framework.

### Control Infrastructure

As shown in Figure 1, our approach can be viewed as two modules within a larger control subsystem. The *Maintenance Module* handles uncertainty about signal identification and channel fidelity. Consistent with the notion of upward completion, the Maintenance Module supports the *Intention Module*, which handles uncertainty about the recognition of user goals from signals. Surrounding both Modules is the *Conversation Control* subsystem which handles uncertainty about the status of the joint activity, Gricean maxims (Grice, 1975), common ground (a shared knowledge base for dialog), and other higher-level dialog events relevant to the joint activity. As represented by the arrows, the Conversation Control subsystem continually exchanges information with both modules and decides where to focus on grounding mutual understanding. The Conversation Control subsystem also adjusts costs or utilities based on records it keeps of conversation level observations, such as the number of questions asked and the number and recency of repair sequences engaged about speech recognition.

As Clark (1996) points out, multiple levels of analysis form a ladder of co-temporal actions ordered with *upward completion* (Clark, 1996). Actions at a given level of dialog are completed bottom up. Furthermore, evidence that one level is complete is also evidence that all levels below it are complete. In other words, evidence flows downward.

The control infrastructure provides an environment for exploring the value of intuitions behind upward the notions of completion and downward evidence. We are investigating procedures and policies for integrating the results of co-temporal inference at multiple levels.

## Representations and Inference Strategies

Every component of the control infrastructure utilizes Bayesian network models to infer probabilities for decision making. The Bayesian networks allow the system to model dependencies between uncertainties, making it possible to reason about sources of misunderstanding in a structured way. Each module consists of at least *one Decision-Based Transition Network* (DTN): a finite automaton with transitional inputs set by decision-theoretic functions. While using a finite automaton to model grounding is not new, a DTN utilizes Bayesian inference and expected value computation to decide what transitions to follow to maximize mutual understanding and minimize collaborative effort. The structure of a DTN encodes conversational strategies for resolving uncertainty.

### Resolving Uncertainty

Grounding in dialog is carried out in what sociolinguists call “adjacency pairs”: an ordered pair of utterances where the second pair part depends on the first by being “conditionally relevant,” *i.e.*, relevant and expectable (Schegloff & Sacks, 1973). The prototype is a question and answer. Adjacency pairs are represented within the structure of a DTN. For example, in handling requests in the Intention Module, the conditionally relevant response is an uptake. However, if there is uncertainty about the request, it may be cost efficient to initiate an embedded “side sequence” (Jefferson, 1972) such as question to repair any misunderstanding. In a DTN, probabilities are provided with Bayesian networks.

In the Intention Module, the primary uncertainty is the goal of the user. Such goals cannot be observed directly so the system infers a probability distribution over possible goals given all the linguistic and contextual, or nonlinguistic, evidence it has observed so far. We have presented details of the machinery employed in the Intention Module in earlier work (Horvitz & Paek, 1999). In the work, we hierarchically decompose the problem of discriminating goals into sets of subgoal discrimination problems at progressively greater levels of detail and represent the problem of discriminating among subgoals at different levels with level-specific Bayesian networks. The approach provides a means for referring to problems of understanding and information gathering at specific levels, and, thus, provides “handles” for grounding a user’s goals. In that work, an approximate decision-theoretic analysis is employed to make decisions about navigation in the hierarchy. If the probability of the most likely goal given the evidence, written $p(\text{Goal}|E)$, does not exceed a probability threshold, determined by a cost—benefit analysis for progression to a conclusion or to a more detailed level of analysis (see Horvitz, 1999 for details on deriving a probability threshold), the system has to decide between two types of conversational strategies, encoded in the DTN as transitional inputs: Inquire_Goal, which explicitly asks the user to confirm the most likely goal, and Ask_Repair, which uses a side sequence to gather more information. Additional probability thresholds determine which transitional input is selected. If $p(\text{Goal}|E)$ does not exceed a certain threshold, Inquire_Goal is ruled out. Furthermore, since failures to reach mutual understanding at the intention level can sometimes be the result of lower level problems, the system also considers uncertainty at the maintenance level in its decision making. In the Maintenance Module, a Bayesian network is used to obtain a probability distribution over three states: CHANNEL AND SIGNAL, CHANNEL AND NO SIGNAL, and NO CHANNEL. A Bayesian network for a Maintenance Module is displayed in Figure 2. The node MAINTENANCE STATUS represent the variables containing the three states at adjacent times.

As before, probability thresholds are used to determine which instantiation of Ask_Repair to use. If $p(\text{Channel and Signal}|E)$ exceeds another threshold, the system asks for an elaboration, which generally suggests an intention level misunderstanding but does not rule out a maintenance level cause such as misparsing.

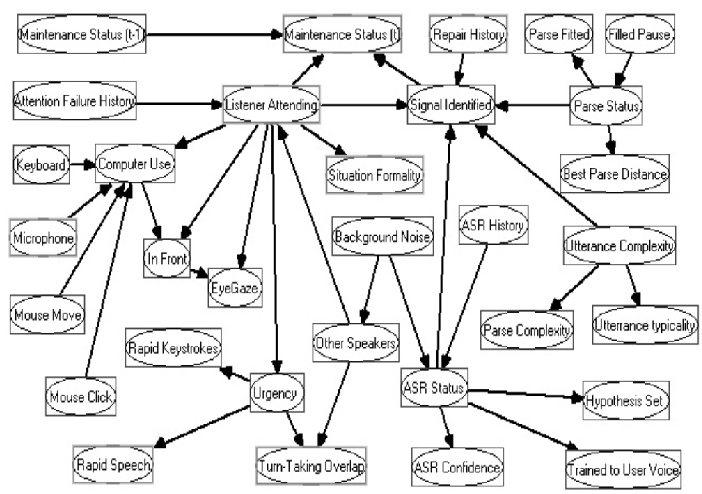


Figure 2. A portion of a temporal Bayesian network for reasoning about misunderstanding in the Maintenance Module.

Another instantiation of Ask_Repair is asking for a repeat. If the most likely maintenance level state is CHANNEL AND NO SIGNAL, then the system will point out that it may be having a signal level problem while requesting a repeat. This acknowledges that resolving a misunderstanding is a collective process since the user is now in a better position to help resolve the problem. The principle underlying this cooperative behavior in dialog has been called the "principle of least collaborative effort (Clark & Wilkes-Gibbs, 1986). For example, if a user asks the *Bayesian Receptionist* to hail a shuttle but the speech recognition only picks out a few words, the system can echo one of the words in a query frame such as "I'm sorry, [shuttle]? Can you repeat that?" where [shuttle] is just a slot for the word. This query tells the user that the speech recognition picked out some words but perhaps not enough to understand the utterance with a high enough level of confidence at the intention level.

One last type of Ask_Repair utilizes the power of value of information (VOI) analysis to identify the best signals to observe in light of the inferred probabilities of different goals. To compute VOI, the system calculates for every signal, the expected utility of the best decision associated with each value the signal may take on. The analysis sums the expected utility for each value, weighted by the probabilities of observing different values should an observation be made (see Horvitz, Breese, & Henrion, 1988 for background and details on computation of VOI). Once it recommends which signal to observe, a query frame for that signal is used. For example, if VOI recommends observing the word "shuttle," the Receptionist may ask the user if the request has anything to do with getting a shuttle. This may be inappropriate at times so context-dependent costs are assigned to VOI recommendations, as will be discussed further in the next section.

In the Intention Module, the transitional input Ask_Repair brings the DTN to a temporary state where, upon admitting the conditionally relevant second pair part, Repair_Admitted, the DTN moves back to its original state. There, the Receptionist performs inference again with additional information acquired from the side sequence. This process of continual refinement through grounding in adjacency pairs is an effective way of clearing up misunderstanding.

### Conversation Control

The Conversation Control subsystem facilitates the sharing of evidence between Modules. Consistent with the notion of downward evidence, once a transitional input is selected in the Intention Module DTN, evidence is sent downwards to fix transitional inputs in the Maintenance Module, as shown in the arrow in Figure 1. For example, when the user provides Repair_Admitted to Ask_Repair, that fixes a transitional input in a Maintenance Module DTN specifying that the question asked was attended to by the user. If the dialog should continue and inference has to be performed on a maintenance level Bayesian model, a variable indicating that the user was attentive to the previous utterance gets instantiated.

As stated previously, transitional inputs in a DTN are set by decision-theoretic functions. In some cases, that involves only fixed probability thresholds. In other cases, dynamically changing costs and benefits (as expressed in utilities) associated with taking a particular action needs to be evaluated. Since the Conversation Control subsystem keeps track of higher-level events, DTNs in both Modules submit utilities for proper alignment with dialog context and social factors. For example, in the Intention Module, the Conversation Control subsystem magnifies the cost of asking a question as the number of questions asked within the same adjacency pair increases. Hence, while the cost of asking one question may be slight, it rapidly increases (perhaps even exponentially) with the number of questions asked about the same request. We are currently exploring utility functions elicited from users in psychological studies for dynamically adjusting cost.

The Conversation Control subsystem is also responsible for discerning when the joint activity has been terminated. It does so by using its own Bayesian network with a distribution over conversation level states such as ADVANCING ACTIVITY and UNEXPECTED TERMINATION.

## Observational and Psychological Studies

To obtain priors and utilities for the Bayesian Receptionist we conducted observational and psychological studies of the receptionist domain. In a related paper (Horvitz & Paek, 1999), we elucidate how detailed interviews and videotapes of people interacting with three receptionists revealed a key set of variables and states relevant to the problem of diagnosing the goals of a user. Variables included not only linguistic signals, as in a parse of the initial utterance, but also visual cues, such as the appearance and trajectory of the user. Knowledge engineering techniques such as eliciting conditional probabilities from the domain experts—receptionists at the Microsoft corporate campus—assisted the assessment of prior distributions for these variables.

Along with variables and their probability distributions, we have been collecting utilities that may be best represented as functions of context. For the Intention Module, we have identified three classes of costs matching three types of observable variables in the Bayesian networks: visual, conceptual, and term features. In recommending a state to observe through VOI, the system breaks down the analysis into the corresponding types and assigns an initial cost.

- **Visual features**: For people, the cost of looking is negligible. For a dialog system, the cost reflects the amount of computational resources necessary to activate and control online visual subsystems.

- **Conceptual features**: The natural language parser, NLPwin (Heidorn, 1999) extracts syntactic, logical, and semantic information from an utterance, which we classify as conceptual features. An example of a semantic feature for the noun "ride" is its association with transportation. To inquire about the concept of transportation with a user may be inappropriate so its initial cost is set high.
- **Term features**: Term features are evocative sets of words and phrases, called *metanyms* (Heckerman and Horvitz, 1998), spotted in an utterance. The cost of asking about whether a particular word is related to the request of the user should be relatively lower than for conceptual features.

After an initial cost is assigned to the three types of features, the Conversation Control subsystem adjusts the cost according to dialog context (based on collected utility functions) and combines it with the utilities of selecting various types of transitional inputs.

Different types and numbers of questions employed in a process of grounding result in different degrees of irritation for the users. In order to obtain utilities for various types of transitional inputs in a DTN, we conducted a questionnaire study at Stanford University about interacting with a receptionist in a hypothetical situation. We asked subjects to imagine scenarios where they would make a typical request to a receptionist, and where the receptionist, who did not understand the request, would respond according to various strategies, such as requesting that the subject repeat the request (e.g., "I'm sorry, can you repeat that?"). Subjects rated each response on a Likert scale. One of the results demonstrate that subjects (n = 104) *significantly* (t = 9.86, $p < .0001$) preferred a simple request for a repeat over a request for an elaboration (e.g., "I'm sorry, I need more information."). One explanation of this result is that the cost of repair implied by an elaboration, which occurs at the intention level, is greater than the cost of repair implied by a repeat.

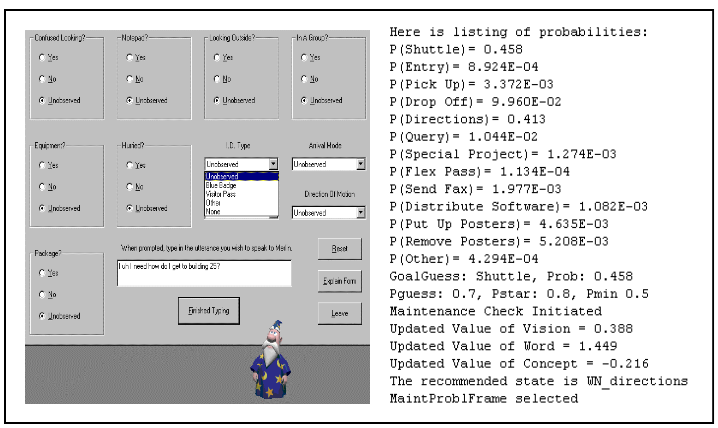


Figure 3. The *Bayesian Receptionist* inferring the goals of the user from an initial utterance and performing value of information analysis at the intention level.

This questionnaire led to the investigation of several hypotheses that we are currently conducting for generating natural responses.

- **Upward Completion Cost Hypothesis**: The cost of repair increases with the level at which failure occurs in the mutual understanding for grounding. Hence, failure and repair at the channel level is preferred to failure repair at the signal level, and so forth.
- **Uncertainty Specification Hypothesis**: Consistent with the Principle of Least Collaborative Effort, interlocutors prefer repair strategies that specify the details about the source of uncertainty.

Additional data and enrichment of contextual issues under consideration will be useful in confirming, modifying, or disconfirming these hypotheses. Psychological studies exploring these two hypotheses have already revealed automated reasoning strategies we did not previously consider. For example, even if the Intention Module may indicate high uncertainty, it may be better for the system to capitalize on a strategy that feigns a maintenance level problem in order to reduce cost; in this case, it may not always be best to be honest about the likely source of failure. From such psychological studies and questionnaires, we are eliciting utilities that can also be used for general decision analysis or for identifying probability thresholds (Horvitz, 1999).

We are also obtaining utilities for transitional inputs that depend on predictive variables such as time. We now present a runtime demonstration of how the *Bayesian Receptionist* deals with misunderstanding.

## Runtime Demonstration

In dialog with a user, the *Bayesian Receptionist* may encounter a misunderstanding that is not restricted to one level of uncertainty. Although the *Bayesian Receptionist* infers the goals of a user and provides the appropriate

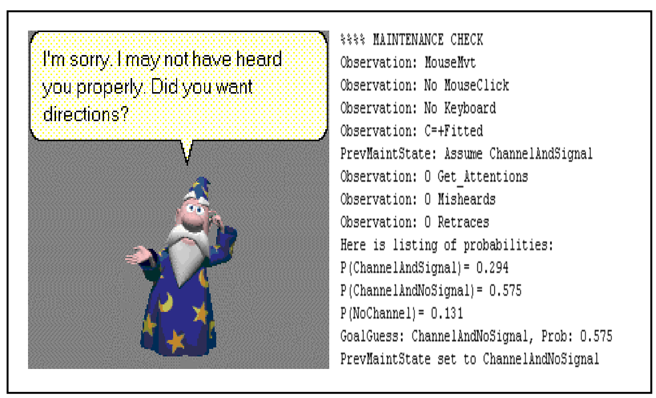


Figure 4. After performing inference at the maintenance level, the *Bayesian Receptionist* decides to ask about a word feature.

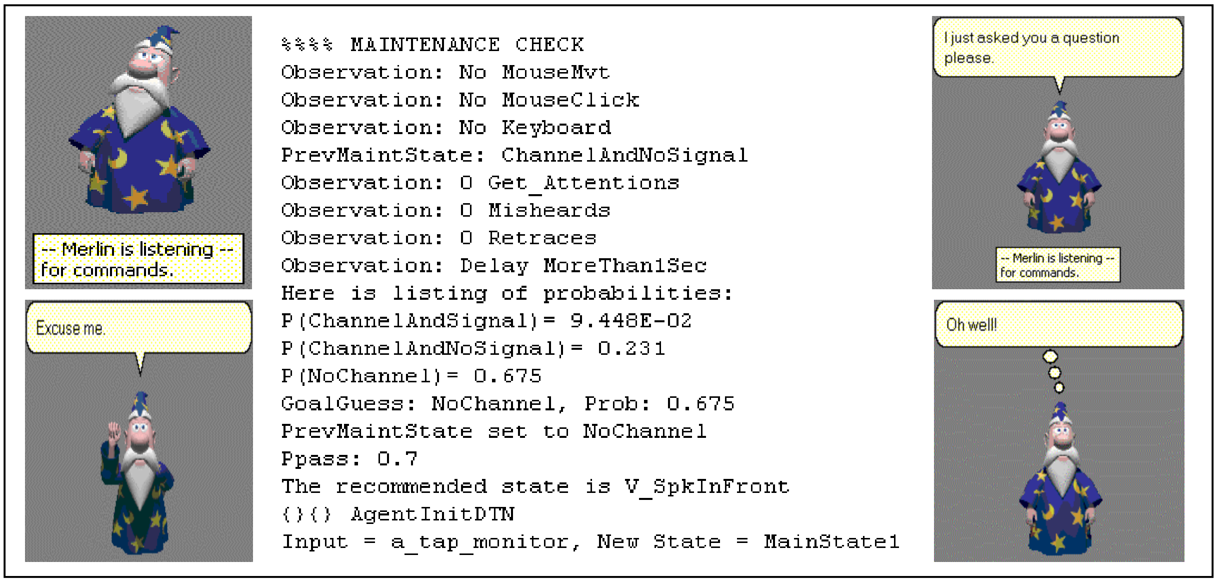


Figure 5. The *Bayesian Receptionist* checks the maintenance level twice to ultimate infer that the joint activity has been unexpectedly terminated.

service in the Intention Module, the system must be able to recognize and resolve uncertainties that cross multiple levels. Figure 3 captures such a situation. A user approaches the *Bayesian Receptionist*, here shown as the user interface agent Merlin, and mutters "I uh …I need …how do I get to building 25?" Notice that the user restarts the utterance twice. Spontaneous speech abounds in restarts and other dysfluencies (Clark, 1994; Stolke & Shriberg, 1996) which may cause problems for speech recognition and natural language parsing. For illustrative purposes, that utterance has been typed into a textbox.

The *Bayesian Receptionist* computes a probability distribution over the goals of the user and find that the most likely goal, a request for a SHUTTLE fails to exceed the probability threshold (Pguess) for Inquire_Goal in the Intention Module DTN. Notice that the second most likely goal, DIRECTIONS, is very close. Since the maximum goal probability is less than a threshold for checking the maintenance level (Pmin), the Receptionist performs inference over a Bayesian network in the Maintenance Module. The results are displayed in the Figure 4.

Here, the most likely state of the maintenance level is CHANNEL AND NO SIGNAL, an apt assessment given a natural language parser that is not equipped to handle restarts.

This information is passed via the Conversation Control subsystem to the Intention Module which now evaluates the costs and benefits of selecting various types of Ask_Repair. Using VOI to consider the best observations to make, drawing from three types of features, and updating the initial costs of these features in the Conversation Control subsystem, the system recommends asking a question that tries to observe the word "directions," or any related metanyms, which would ideally discriminate between between the goals of SHUTTLE and DIRECTIONS. Since the most likely maintenance state is CHANNEL AND NO SIGNAL, the system chooses a query frame that specifies a possible misunderstanding at the maintenance level. The output is the combination, "I'm sorry, I may not have heard you properly. Did you want directions?"

Imagine the situation where, in the midst of such an interaction with the *Bayesian Receptionist*, a user notices that a shuttle has just approached the front of the building and decides to simply break off the interaction. A DTN in the Maintenance Module is expecting a conditionally relevant second pair part to the question, but the user times out of the listening mode, as shown in Figure 5. This is very unusual for a dialog so the *Bayesian Receptionist* performs inference over the maintenance level Bayesian network again and this time identifies NO CHANNEL as the most likely state. Since the maximum probability does not exceed a threshold for checking if the activity has been completed, given multiple observations, including the observation that the user was just previously attentive, the system decides to perform a side sequence in the maintenance DTN and "knocks" on the display glass. The user does not respond again for 10 seconds. This gets instantiated in the second round of inference and the most likely state is now NO CHANNEL. The Conversation Control subsystem considers this evidence and performs inference on its own Bayesian network to check the status of the joint activity. The most likely status there is UNKNOWN TERMINATION of the activity so the *Bayesian Receptionist* just sighs.

## Conclusion and Future Directions

We have outlined four levels of uncertainty for grounding mutual understanding through the use of Bayesian networks and DTNs within modules and a larger control subsystem. In contrast to the majority of automated dialog systems, which focus only on the intention level, we have described an infrastructure that recognizes that failures in dialog can happen at any of the levels described. Rather than writing ad hoc policies to meet such failures, we describe explicit representations, inference strategies, and decision procedures to reason about and repair misunderstanding at any level of analysis. The long-term payoff we hope will be systems that handle uncertainty and misunderstanding in a natural manner, and that are ultimately more robust to failure. We foresee that the approach of treating dialog as a joint activity will change the experience of interacting with computers in fundamental ways. Our research is continuing in the realm of building rich models for providing the machinery for grounding in conversation. We are pursuing more expressive probabilistic models for diagnosing failures and for making expected-utility decisions about the best actions to take to address potential problems in conversation. Research in this area must continue to rely on psychological studies to identify the perceived costs of conversational states and actions.

## References


Brennan, S.A. 1998. The grounding problem in conversations with and through computers. In *Social and Cognitive Approaches to Interpersonal Communication,* 201-225. Lawrence Earlbaum Associates

Brennan, S.A. & Hulteen, E. 1995. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems* 8: 143-151.

Clark, H.H. 1994. Managing problems in speaking. *Speech Communication* 15: 243-250.

Clark, H.H. 1996. *Using Language*. Cambridge University Press.

Clark, H.H. & Brennan, S.A.. 1991. Grounding in communication. In *Perspectives on Socially Shared Cognition*, 127-149. APA Books

Clark, H.H. & Schaefer, E.F. 1987. Collaborating on contributions to conversations. *Language and Cognitive Processes* 2(1): 19-41.

Clark, H.H. & Schaefer, E.F. 1989. Contributing to discourse. *Cognitive Science* 13: 259-294.

Clark, H.H. & Wilkes-Gibbs, D. 1990. Referring as a collaborative process. In *Intentions in Communication*, 463-493. MIT Press.

Cohen, P.R. & Levesque, H.J. 1991. Teamwork. *Nous*, 25(4): 487-512.

Cohen, P.R. and Levesque, H.J. 1994. Preliminaries to a collaborative model of dialogue. *Speech Communication* 15: 265-274.

Conati, C., Gertner, A., Van Lehn, K., and Druzdzel, M., 1997. Online student modeling for coached problem solving using Bayesian networks. *Proc. of the Sixth International Conference on User Modeling*, 231-242. Springer-Verlag.

Dillenbourg, P., Traum, D. & Schneider, D. 1996. Grounding in multi-modal task-oriented collaboration. *Proc. of the EuroAI & Education Conference*.

Edmonds, P.G. 1993. A computational model of collaboration on reference in direction-giving dialogues, M.Sc. thesis, Dept. of Computer Science, University of Toronto.

Goodwin, C. 1986. Between and within: Alternative sequential treatments of continuers and assessments. *Human Studies* 9:205-217.

Grice, H.P. 1957. Meaning. *Philosophical Review* 66: 377-388.

Grice, H.P. 1975. Logic and conversation. In *Syntax and Semantics 3: Speech Acts*, 41-58. Seminar Press.

Grosz, B.J. & Sidner, C.L. 1990. Plans for discourse. In *Intentions in Communication*, 417-444. MIT Press.

Haddadi, A. 1995. *Communication and Cooperation in Agent Systems: A Pragmatic Theory*. Springer-Verlag.

Heeman, P. 1991. A computational model of collaboration on referring expressions. M.Sc. thesis, Dept. of Computer Science, University of Toronto.

Heeman, P., and Hirst, G. 1992. Collaborating on referring expressions (TR 435). Department of Computer Science, University of Rochester, Rochester NY.

Heidorn, G. 1999. Intelligent writing assistance. In *A Handbook of Natural Language Processing Techniques*. Marcel Dekker.

Hirst, G., McRoy, S., Heeman, P., Edmonds, P., & Horton, D. 1994. Repairing conversational misunderstandings and non-understandings. *Speech Communication* 15: 213-229.

Horvitz, E. 1997. Agents with beliefs: Reflections on Bayesian methods for user modeling. *Proc. of the Sixth International Conference on User Modeling*, 441-442. Springer-Verlag.

Horvitz, E. 1999. Principles of mixed-initiative user interfaces. *Proc. of CHI '99, ACM SIGCHI Conference on Human Factors in Computing Systems*, Pittsburgh, PA, May 1999, 159-166. ACM Press.

Horvitz, E. & Paek, T. 1999. A computational architecture for conversation. *Proc. of the Seventh International Conference on User Modeling*, 201-210. Springer Wien.

Horvitz, E., Breese, J., and Henrion, M., 1988. Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2: 247-30.

Horvitz, E., Breese, J., Heckerman, D., Hovel, D., Rommelse, D., 1998. The Lumiere project: Bayesian user modeling for inferring the goals and needs of software users. *Fourteenth Conference on Uncertainty in Artificial Intelligence,* 256-265. Morgan Kaufmann.

Jefferson, G. 1972. Side sequences. In D. Sudnow, ed., *Studies in Social Interaction*, 294-338. Free Press.

Levinson, S.C. 1992. Activity types and language. In P. Drew and J. Heritage, eds., *Talk at Work*, 66-100. Cambridge University Press.

Schegloff, E.A. & Sacks, H.. 1973. Opening up closings. *Semiotica* 8:289-327.

Stolcke, A. & Shriberg, E.E. (1996). Statistical language modeling for speech disfluencies. *Proc. International Conference on Acoustics, Speech and Signal Processing*, 405-408.

Suchman, L.A. 1987. *Plans and Situated Actions: The Problem of Human/Machine Communication*. Cambridge University Press.

Traum, D. 1994. A computational theory of grounding in natural language conversation, Ph.D. diss., Dept. of Computer Science, Rochester.

Traum, D. & Dillenbourg, P. 1996. Miscommunication in multi-modal collaboration. *AAAI Workshop on Detecting, Repairing, And Preventing Human--Machine Miscommunication*, 37-46.

Traum, D. & Dillenbourg, P. 1998. Towards a normative model of grounding in collaboration,. *ESSLLI-98 Workshop on Mutual Knowledge, Common Ground and Public Information*.